



Duplication de génome et évolution de la famille Sox chez les poissons téléostéens

Emilien Voldoire

► To cite this version:

Emilien Voldoire. Duplication de génome et évolution de la famille Sox chez les poissons téléostéens. Biologie animale. Ecole normale supérieure de lyon - ENS LYON, 2013. Français. NNT : 2013ENSL0875 . tel-01124192

HAL Id: tel-01124192

<https://theses.hal.science/tel-01124192>

Submitted on 6 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

en vue de l'obtention du grade de

Docteur de l'Université de Lyon, délivré par l'École Normale Supérieure de Lyon

Sciences de la Vie

Institut de génomique fonctionnelle de Lyon

École Doctorale de Biologie Moléculaire, Intégrative et Cellulaire

présentée et soutenue publiquement le 17 décembre 2013

par Monsieur Emilien VOLDOIRE

Duplication de génome et évolution de la famille Sox chez les poissons téléostéens

Directrice de thèse : Madame Delphine GALIANA

Après l'avis de : Monsieur Didier CASANE

Monsieur Marc ROBINSON-RECHAVI

Devant la commission d'examen formée de :

<i>Monsieur Didier CASANE</i>	<i>Rapporteur</i>
-------------------------------	-------------------

<i>Madame Delphine GALIANA</i>	<i>Directrice</i>
--------------------------------	-------------------

<i>Madame Élisabeth HUGUET</i>	<i>Membre</i>
--------------------------------	---------------

<i>Monsieur Marc ROBINSON-RECHAVI</i>	<i>Rapporteur</i>
---------------------------------------	-------------------

<i>Madame Marie SÉMON</i>	<i>Membre</i>
---------------------------	---------------

<i>Monsieur Jean-Nicolas VOLFF</i>	<i>Président</i>
------------------------------------	------------------

SOMMAIRE

INTRODUCTION	5
Phylogénie et diversité des animaux vertébrés	7
Introduction à la phylogénie des chordés	7
Les vertébrés ou l'acquisition d'une tête	9
Le succès évolutif des poissons téléostéens	11
Quel matériel génétique pour l'innovation ?	17
Comment dupliquer un gène ?	19
Duplication de gène par duplication segmentaire	19
Duplication de gène par duplication de génome	20
Polyploïdie et duplication de génome	23
Un cas extrême de mutation : la polyploïdisation ou la duplication complète d'un génome	23
Polyploïdie : une impasse évolutive ?	24
Comment identifier les génomes paléopolyploïdes ?	25
Paléopolyploïdie chez les vertébrés	26
Quel devenir pour les génomes et les gènes dupliqués ?	31
Le mystère de la diploïdisation	31
Un gène ou deux ?	32
Les mécanismes de préservation des gènes dupliqués	33
La co-optation ou l'émergence de nouvelles fonctions	38
La portée évolutive des duplications de génome	41
Duplications de génome, origine et diversification des vertébrés	41
Duplication de génome et spéciation	43
Duplication de génome, exaptation et complexification	46
La famille des facteurs de transcription Sox	47
Découverte du premier membre de la famille Sox	47
Évolution et diversité fonctionnelle de la famille Sox	49

RÉSULTATS _____ 53

Partie I. Expansion et évolution de la famille Sox chez les poissons téléostéens _____ 55

Partie II. Duplication de génome et évolution lignée-spécifique : *sox11* chez le poisson-zèbre et le platy _____ 88

Divergence lignée-spécifique des paralogues *sox11* après duplication de génome 89

Perspectives 105

DISCUSSION _____ 109

Duplication de génome et expansion de la famille Sox 111

Devenir des gènes *sox* dupliqués chez les téléostéens 114

Expansion de la famille Sox : diversification fonctionnelle ? 116

Duplication de la famille Sox : impact sur la spéciation ? 120

REFERENCES _____ 125

REMERCIEMENTS _____ 135

ANNEXE _____ 137

The Sox family: evolution, biology and its role in diseases 138

TABLE DES FIGURES

Figure 1: Phylogénie simplifiée des deutérostomiens.	6
Figure 2 : Phylogénie simplifiée des vertébrés.	8
Figure 3: Exemples de diversité écologique chez les poissons téléostéens.	10
Figure 4: Exemples de diversité morphologique chez les poissons téléostéens.	10
Figure 5: Phylogénie des poissons actinoptérygiens calibrée dans le temps sur la base de marqueurs moléculaires et de 36 fossiles (d'après Near et al. 2012).	12
Figure 6: Mécanismes génétiques principaux qui conduisent à l'émergence de nouveaux gènes.	16
Figure 7: Génomes diploïde et polyploïde (Otto 2007).	22
Figure 8: Matrice de gènes homologues (Van de Peer 2004).	26
Figure 9: Scénario évolutif des complexes Hox chez les chordés.	28
Figure 10: Diploïdisation (Otto 2007).	30
Figure 11: Duplication de génome et effet de dosage.	34
Figure 12: Mécanismes de préservation des gènes dupliqués.	36
Figure 13: Voies de biosynthèse de la valine et de l'isoleucine chez la levure <i>Saccharomyces cerevisiae</i> (Conant & Wolfe 2008).	40
Figure 14: Mécanismes génétiques d'isolement reproductif.	44
Figure 15: Classification des membres de la famille Sox	48
Figure 16: Phylogénie de la famille Sox.	50
Figure 17: Scénario pour l'expansion de la famille Sox chez les bilatériens.	112
Figure 18: Expression des gènes <i>sox</i> chez quatre vertébrés adultes.	118
Figure 19: Patrons de perte des gènes <i>sox</i> après duplication.	121
Figure 20: Non-fonctionnalisation lignée-spécifique après duplication.	123

INTRODUCTION

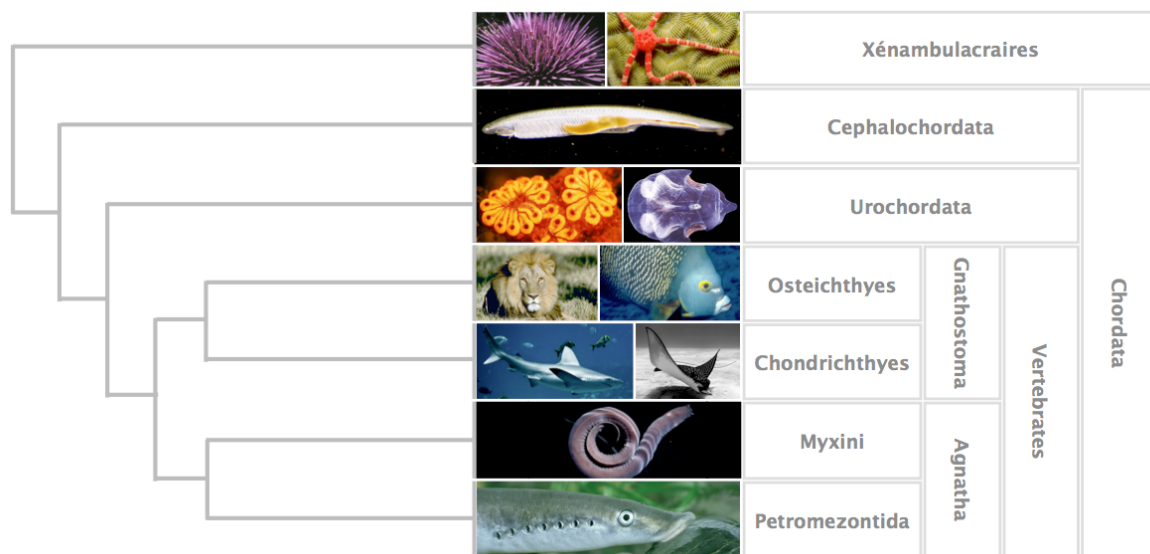


Figure 1: Phylogénie simplifiée des deutérostomiens. Les xénambulacraires, représentés ici par un oursin et une étoile de mer, constituent le groupe frère des chordés (qui forment ensemble le groupe des deutérostomiens). Les urochordés sont quant à eux le groupe frère des vertébrés (ou crâniates). Les Osteichthyes sont illustrés sur cette phylogénie par un mammifère (le lion) et un poisson téléostéens (le poisson-ange français), qui appartiennent respectivement aux deux grands groupes de ce clade : sarcoptérygiens et actinoptérygiens.

Phylogénie et diversité des animaux vertébrés

L'arbre de la vie est foisonnant, et la diversité du vivant est flagrante. L'étude de la structure de l'arbre du vivant (les eucaryotes en particulier) pour comprendre l'organisation de la diversité des espèces, et inférer les mécanismes à l'origine de la biodiversité, est actuellement portée par la révolution génomique. En effet, d'une part, les données de génomes contribuent à améliorer nos connaissances sur la structure de la diversité, grâce à la reconstruction des relations de parenté entre espèces à partir de très nombreux marqueurs moléculaires (Lecointre *et al.* 2010). Par exemple, la reconstruction d'arbres phylogénomiques (reconstruction phylogénétique à partir d'un très grand nombre de gènes) est à l'origine d'une redéfinition des relations de parenté de nombreux groupes d'animaux, notamment chez les poissons téléostéens (présentés ci-dessous). D'autre part, le séquençage en masse des génomes de différentes espèces fournit une opportunité sans précédent pour étudier la dynamique évolutive des génomes, et ce par une approche de génomique comparative. Ce premier chapitre vise à présenter succinctement la phylogénie et la diversité des animaux vertébrés, en particulier du groupe des téléostéens. Les autres chapitres de cette introduction sont destinés à exposer les mécanismes génétiques proposés dans la littérature, en particulier le rôle des duplications de gènes et de génome, pour expliquer la biodiversité des espèces vertébrés.

Introduction à la phylogénie des chordés

Les chordés représentent, avec les xénambulacraires (échinodermes, hémichordés et xénoturbellides), le second grand groupe des deutérostomiens (figure 1). Les chordés doivent leur nom à une structure embryonnaire nommée «chorde» issue du mésoderme axial. Cette structure est impliquée dans l'établissement de l'axe antéro-postérieur de l'embryon et dans la différenciation de plusieurs autres structures uniques aux chordés (Stemple 2005). Les chordés sont constitués de trois sous-groupes d'animaux existants : les céphalochordés (~25 espèces), les urochordés (~1500 espèces) et les vertébrés (ou crâniates, ~50000 espèces). Le séquençage de nombreux génomes de chordés, et le développement de la phylogénomique est à l'origine d'une redéfinition récente des relations de parenté entre ces trois grands groupes (Delsuc *et al.* 2006). Ainsi, la phylogénie moléculaire des chordés montre que les urochordés sont les plus proches parents des vertébrés et constituent ensemble un groupe

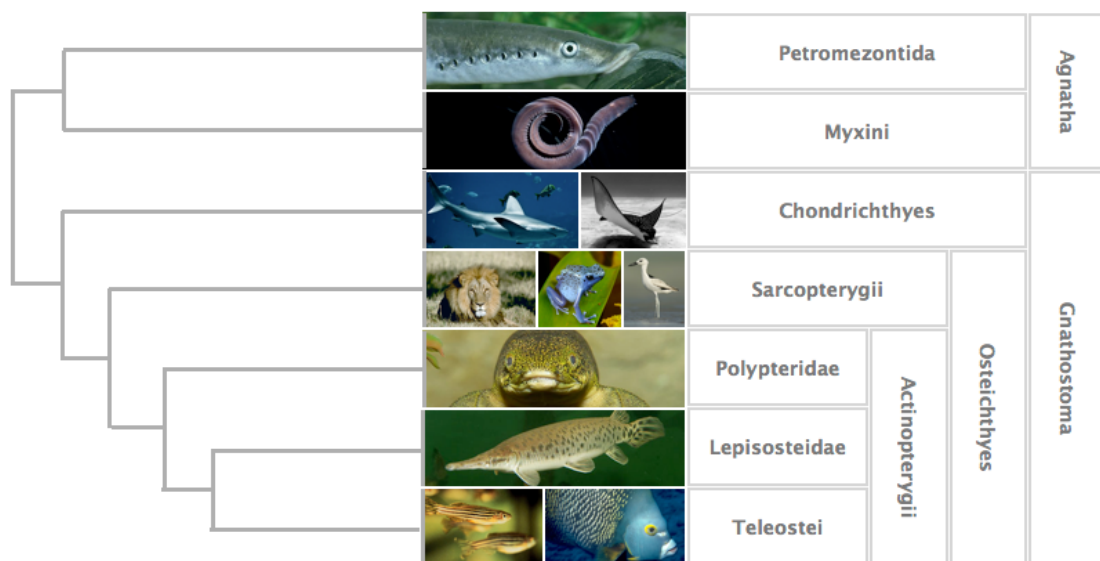


Figure 2 : Phylogénie simplifiée des vertébrés. Les vertébrés sont divisés en deux sous-groupes : les gnathostomes et les agnathes. Les sarcoptérygiens, groupe frère des actinoptérygiens, regroupent les tétrapodes (mammifères, amphibiens et sauropsides) et les cœlacanthes et dipneustes (non représentés).

monophylétique nommé olfactores. Les céphalochordés, qui présentent des caractéristiques morphologiques similaires aux vertébrés, étaient considérés initialement comme leurs plus proches parents. Finalement, des études récentes montrent que les urochordés possèdent certaines structures embryonnaires à l'état rudimentaire (la crête neurale notamment), absentes des céphalochordés, qui étaient considérées jusqu'alors comme uniques aux vertébrés (Mazet *et al.* 2005, Holland 2009, Abitua *et al.* 2012). Ces nouvelles données confortent la monophylie du groupe des olfactores qui partagent ainsi plusieurs caractères synapomorphiques (Cañestro *et al.* 2013).

Les vertébrés ou l'acquisition d'une tête

Les vertébrés, avec près de cinquante mille espèces, constituent de loin le groupe le plus représenté chez les chordés. Ils se divisent en deux sous-groupes définis par la présence ou non d'une mâchoire : les gnathostomes (*gnathos*, mâchoire, et *stoma*, bouche), qui regroupent la grande majorité des vertébrés, et les agnathes, qui comprennent les lamproies et les myxines (~60 espèces) (figure 2). Les gnathostomes sont quant à eux divisés en osteichthyes (*osteon*, os, et *ichthys*, poisson), représentés par les actinoptérygiens (poissons à nageoires rayonnées) et les sarcoptérygiens (cœlacanthes, dipneustes, et tétrapodes), et en chondrichthyes (*chondros*, cartilage) qui comprend environ mille espèces. Il faut noter ici que les poissons, qui sont répartis dans les quatre taxons agnathes, chondrichthyes, actinoptérygiens, et sarcoptérygiens constituent un groupe paraphylétique qui désigne l'ensemble des vertébrés non-tétrapodes. Les actinoptérygiens, qui incluent les poissons téléostéens présentés ci-après, regroupent l'écrasante majorité des espèces de poissons actuels. Les animaux vertébrés sont habituellement définis par des caractères anatomiques tels que les vertèbres, le crâne, ou encore les tissus dentaires. Cependant, ces caractères ne sont pas communs à l'ensemble des animaux vertébrés et/ou sont parfois difficiles à définir. Actuellement, les biologistes s'accordent pour définir les vertébrés selon trois principaux caractères embryonnaires : la crête neurale, les placodes neurogéniques et un cerveau tripartite (Holland 2009). Ces structures embryonnaires, présentes à l'état rudimentaire chez les urochordés, sont chez les vertébrés à l'origine de nombreux types cellulaires qui constituent des innovations évolutives uniques aux vertébrés (os, cartilage, placodes et nerfs sensoriels notamment). Ensemble, ces structures embryonnaires participent entre autres chez les vertébrés au développement d'une tête et d'un cerveau pourvu d'organes



Figure 3: Exemples de diversité écologique chez les poissons téléostéens. Le saumon royal (en haut à gauche) est rencontré dans les torrents de montagne, tandis que le poisson-dragon, poisson bioluminescent, vit dans les abysses (en bas à gauche). Les poissons des glaces (famille des Channichthyidae) vivent en antarctique (en haut à droite) tandis que la mer des caraïbes abrite la rascasse volante (en bas à droite).

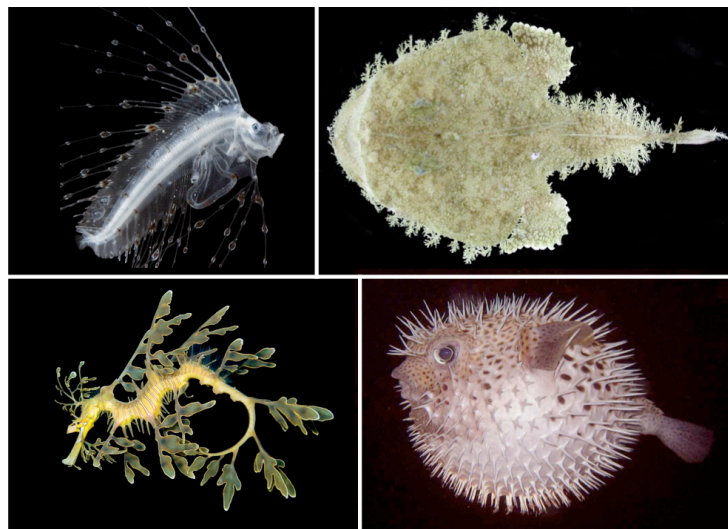


Figure 4: Exemples de diversité morphologique chez les poissons téléostéens. Le roi des harengs (en haut à gauche) peut mesurer jusqu'à 5 mètres de long à l'âge adulte. La baudroie réticulée en haut à droite a la forme d'une crêpe. L'hippocampe feuille ressemble à une algue (en bas à gauche), tandis que le poisson-hérisson (en bas à droite) a la forme d'une boule.

des sens, et sont certainement responsables de la transition vers un comportement de prédation des animaux vertébrés (Gans & Northcutt 1983). Les rares fossiles de vertébrés anciens (datant d'environ 535 Ma) identifiés dans la région de Chengjiang en Chine, qui possèdent un squelette viscéral et des placodes sensorielles putatives, indiquent une origine des vertébrés datant du cambrien inférieur (Holland & Chen 2001). Depuis à présent quelques décennies, l'hypothèse d'un bouleversement d'ordre génétique est proposée pour expliquer l'émergence du plan d'organisation des vertébrés, et les nombreuses innovations morphologiques associées (Ohno 1970, Van de Peer *et al.* 2009). En effet, deux duplications de génome complet sont survenues à la base des vertébrés, et ces événements majeurs ont probablement facilité la diversification et le raffinement des structures embryonnaires des vertébrés. Finalement, les vertébrés constituent non seulement le groupe des chordés le plus important en nombre d'espèces, mais également un groupe d'espèces avec une très grande diversité anatomique, morphologique et écologique.

Le succès évolutif des poissons téléostéens

L'immense majorité des actinoptérygiens est des poissons téléostéens. Les quelques dizaines d'espèces de poissons actinoptérygiens qui ne sont pas des téléostéens sont répartis dans les trois ordres polyptériformes (~10 espèces), acipensériformes (26 espèces) et lépisostéiformes (7 espèces, dont le lépisosté tacheté) et vivent uniquement dans les eaux douces. Le groupe des téléostéens, qui trouve son origine au carbonifère (~300 Ma, Near *et al.* 2012), compte en revanche actuellement plus de 30000 espèces répertoriées (www.fishbase.org), soit autant d'espèces que tous les autres vertébrés réunis (47% des vertébrés sont des téléostéens ; Lecointre *et al.* 2010). En outre, par sa diversité morphologique, biologique et écologique, ce groupe représente un remarquable succès évolutif. Les poissons téléostéens sont de toutes tailles, de toutes formes, de toutes couleurs, et vivent dans tous les milieux aquatiques, marins ou dulçaquicoles, depuis les zones abyssales aux torrents de hautes montagnes, et sous toutes latitudes (Lecointre & Le Guyader 2013 ; figures 3 et 4). Malgré la très grande diversité des téléostéens, trois caractères morphologiques synapomorphiques sont identifiables: la présence d'écaillés élasmoïdes (ganoïdes chez les autres poissons actinoptérygiens), une double mobilité des pièces osseuses externes supérieures de la mâchoire (maxillaire et prémaxillaire), et une nageoire caudale symétrique (le lobe inférieur est aussi développé que le supérieur).

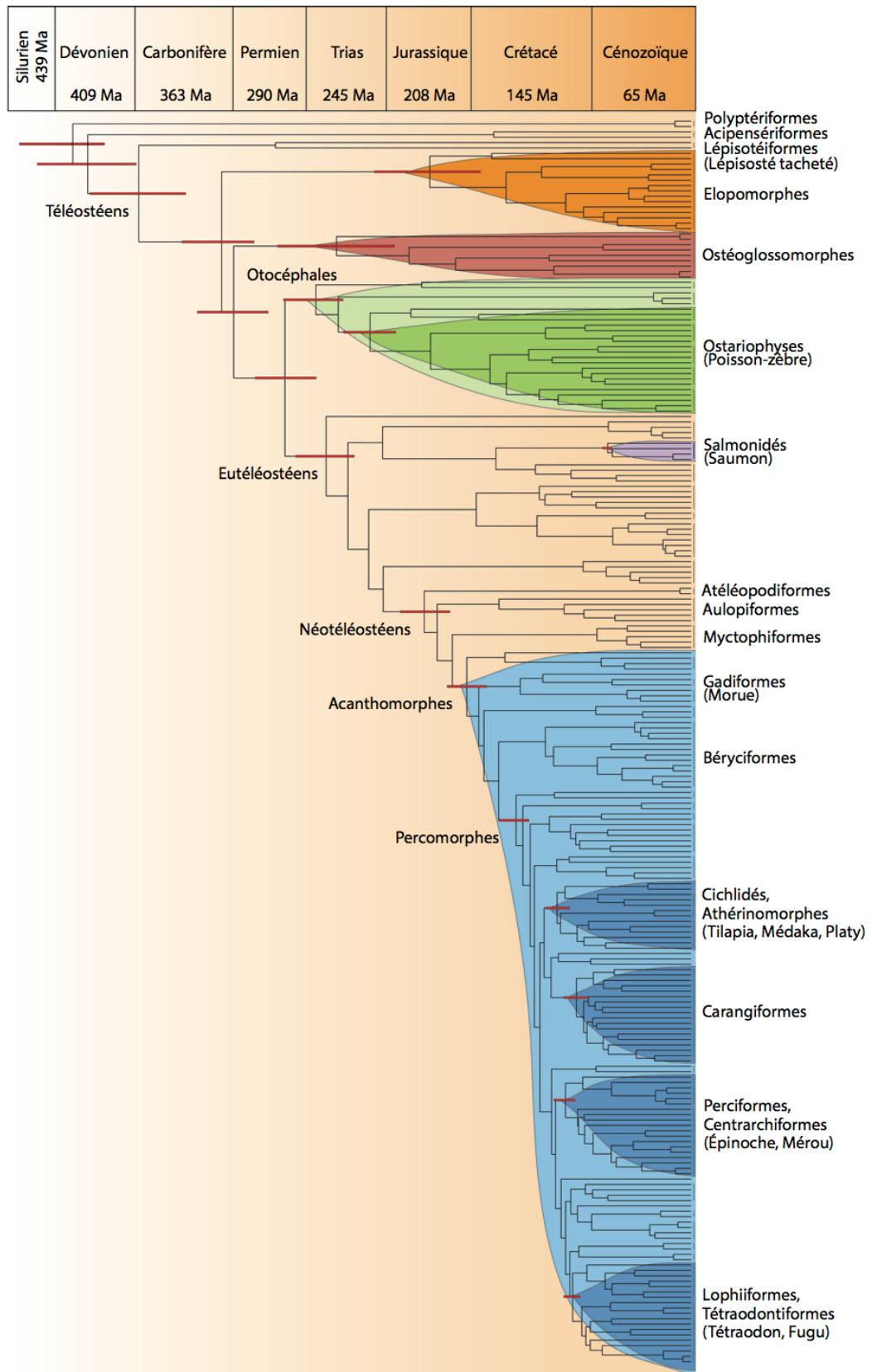


Figure 5: Phylogénie des poissons actinoptérygiens calibrée dans le temps sur la base de marqueurs moléculaires et de 36 fossiles (d'après Near *et al.* 2012). Les barres représentent l'estimation de la date de divergence entre deux lignées. Les espèces les plus étudiées dans ce manuscrit sont indiquées entre parenthèses.

Introduction à la phylogénie des téléostéens

La classification des téléostéens a énormément progressé depuis le début des années 1990. L'accumulation des données moléculaires et la construction de nombreuses phylogénies moléculaires (fondées sur la comparaison de gènes nucléaires et/ou mitochondriaux) sont à l'origine des multiples réarrangements de l'arbre phylogénétique des téléostéens survenus ces dernières années (Lecointre *et al.* 2010, Lecointre & Le Guyader 2013, Near *et al.* 2012). Au niveau de cet arbre phylogénétique (figure 5), les élopomorphes et les ostéoglossomorphes constituent les deux groupes les plus anciens des téléostéens (~1100 espèces). Les autres poissons téléostéens sont divisés en deux grands groupes: les otocéphales et les euté téléostéens, qui ont divergé au permien il y a environ 260 Ma. Le groupe des otocéphales, qui comprend actuellement près de 10000 espèces, représente la part la plus grande de la diversité téléostéenne dulçaquicole. Sur le plan anatomique, toutes ces espèces partagent notamment une connexion entre l'oreille interne et la vessie natatoire, et des similarités dans le squelette de la queue (Lecointre *et al.* 2010). Ce groupe est majoritairement représenté par les ostariophyses (~9000 espèces), qui contient les cypriniformes (dont le célèbre poisson-zèbre), les characiformes (piranhas, néons) et les siluriformes (poisson-chat). Le second grand groupe, les euté téléostéens, est défini anatomiquement par des petits cartilages supplémentaires dans le squelette de la queue, et la présence de l'os stégural (Lecointre *et al.* 2010). Il comprend notamment les protacanthoptérygiens (dont les salmonidés, ~370 espèces) et le groupe des néotéléostéens. Les trois premières lignées appartenant aux néotéléostéens (aulopiformes, atéléopodiformes et myctophiformes) sont toutes composées d'animaux des profondeurs marines, et contiennent pour la plupart des espèces aux morphologies étonnantes : bouche et/ou dents démesurées, organes lumineux (photophores), ou encore yeux télescopiques. Les acanthomorphes complètent le grand groupe des néotéléostéens, et partagent avec les quatre autres lignées un muscle (le rétracteur dorsal) qui relie les éléments dorsaux du squelette branchial aux six premières vertèbres.

Le groupe des acanthomorphes ou la radiation des poissons téléostéens

La diversité des acanthomorphes est considérable: environ 60% des poissons téléostéens sont des acanthomorphes, ce qui représente plus de 17000 espèces. D'un point de vue anatomique, tous les acanthomorphes possèdent des rayons épineux non segmentés dans la première nageoire dorsale et la nageoire anale. Le groupe des acanthomorphes trouve son origine évolutive au crétacé il y a 130 millions d'années

environ et constitue un cas de radiation évolutive. En effet, ce groupe très récent dans l'histoire évolutive des vertébrés représente néanmoins environ 25% des espèces de vertébrés, et ces nombreuses espèces sont caractérisées par une très grande diversité morphologique et écologique. Ainsi, les acanthomorphes ont des formes extrêmement variées qui vont du guppy au thon, en passant par les hippocampes, le rombou et le dragonnet lyre, et vivent dans tous les milieux, des fond abyssaux (lophiiformes notamment) aux torrents des montagnes, et sous tous les climats (Lecointre & Le Guyader 2013). L'avènement des phylogénies moléculaires a profité tout particulièrement à ce large groupe, notamment pour reconstruire l'histoire évolutive de nombreuses espèces qui ne possèdent pas de caractères anatomiques ou morphologiques bien distinctifs (Lecointre *et al.* 2010). Les relations de parenté entre les différents taxons de ce grand groupe sont toujours à l'étude, et la présentation simplifiée de ces relations proposée dans le présent manuscrit (quelques grands groupes, figure 5) se base sur un travail réalisé par l'équipe américaine de Leo Smith en 2012 (à partir de neuf gènes nucléaires ; Near *et al.* 2012). Ainsi, ces travaux confirment que les gadiformes (morue, grenadier) et les beryciformes (béryx, poissons-soldats) constituent deux ordres anciens des acanthomorphes. Les autres acanthomorphes sont rassemblés dans le grand groupe des percomorphes, au sein duquel émergent quatre groupes majeurs encore non nommés, établis principalement sur la base de caractères moléculaires. Le premier représente l'ensemble des cichlidés (tilapia) et des athérinomorphes (composé du célèbre médaka, mais également du platy et de l'orphie). Le second groupe contient les carangiformes, représentés par de très nombreuses espèces, dont l'ensemble des poissons plats (limande). Le troisième grand groupe des percomorphes, qui trouve son origine évolutive il y environ 70 Ma, contient les perciformes (~10000 espèces dont l'épinoche, le dragonnet lyre et le mérou) et les centrarchiformes (perche-soleil). Enfin, les lophiiformes (baudroie) et les tétraodontiformes (tétrodon, fugu) composent un quatrième groupe également récent (~70 Ma). Comme mentionné précédemment, les relations de parenté entre les poissons téléostéens, et particulièrement les acanthomorphes, sont encore à l'étude. Cependant, les travaux réalisés par trois équipes indépendantes (Miya au Japon, Smith aux États-Unis d'Amérique, et Lecointre en France) qui examinent différents gènes (nucléaires et/ou mitochondriaux), indiquent la monophylie des quelques groupes majeurs présentés dans ce manuscrit (Lecointre & Le Guyader 2013).

Dans ce premier chapitre, la brève présentation des animaux vertébrés, et en particulier des poissons téléostéens, souligne la diversité anatomique, morphologique et écologique des espèces de vertébrés. La compréhension des mécanismes à l'origine de la diversité biologique des organismes vivants nécessite de comprendre comment les organismes fonctionnent et se diversifient et comment ils interagissent avec les autres organismes et leur environnement. En particulier, quels sont les mécanismes génétiques qui favorisent l'émergence de nouveaux traits morphologiques, physiologiques et comportementaux ?

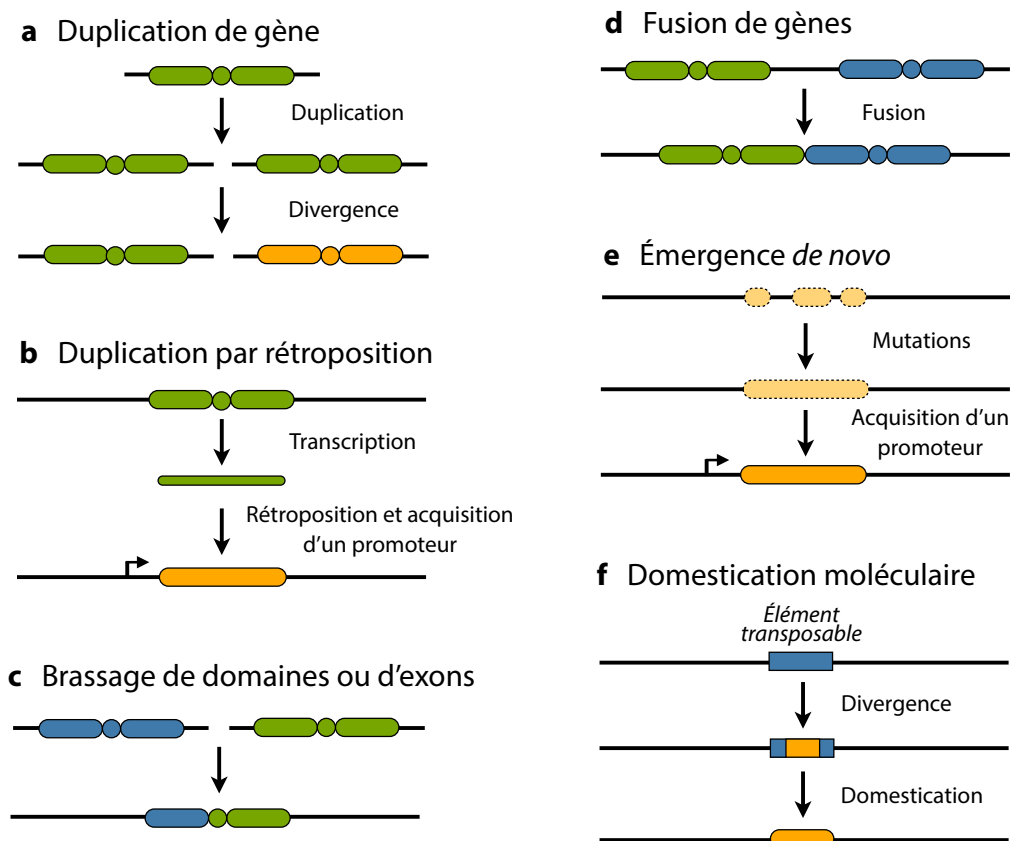


Figure 6: Mécanismes génétiques principaux qui conduisent à l'émergence de nouveaux gènes. Les exons sont représentés par des cylindres et les introns par des cercles. La duplication de gène avec un intermédiaire ADN (depuis la duplication d'un fragment de gène à la duplication de génome) est le mécanisme le plus connu pour l'émergence de nouveaux gènes (a). La duplication avec un intermédiaire ARN nécessite la présence des enzymes de rétroposition dans le cytoplasme et l'acquisition d'un promoteur (b). Le brassage de domaines (ou d'exons) de différents gènes (c) ainsi que la fusion de gènes (d) peut produire des nouveaux gènes chimères. A partir d'une séquence non-codante, des mutations peuvent éliminer les interruptions d'un cadre de lecture potentiel, et suite à l'acquisition d'un promoteur, conduire à l'émergence d'un gène *de novo* (e). La domestication moléculaire (ici d'un élément transposable) peut conduire à de nouveaux gènes (f).

Quel matériel génétique pour l'innovation ?

Comment apparaissent les innovations génétiques ? Depuis l'ère pré-moléculaire de la génétique, les biologistes attribuent un rôle majeur aux nouveaux gènes dans la diversification phénotypique des lignées et des espèces (Haldane 1933, Muller 1935, Chen *et al.* 2013). Au sein des génomes eucaryotes, différents mécanismes conduisent à l'émergence de nouveaux gènes (figure 6, Kaessmann 2010). Ainsi, la duplication de gènes (intermédiaire ADN ou ARN, voir chapitre suivant), la fusion de gènes (gène chimère), et la domestication de gènes (génome parasite par exemple) constituent des mécanismes qui conduisent à de nouveaux gènes à partir de gènes pré-existants. L'émergence de gènes *de novo* (à partir d'une séquence non-codante dans le génome), bien que longtemps considérée comme très peu probable, est également un mécanisme important pour la production de nouveaux gènes (Kaessmann 2010). Par exemple, la famille de gènes *morpheus* est certainement apparue suite à l'émergence *de novo* d'un nouveau gène chez l'ancêtre commun des cercopithécidés (primates catarrhiniens, qui incluent les hominoïdés), dans la mesure où aucune séquence orthologue correspondante n'est détectée en dehors de ce clade (Johnson *et al.* 2001). Il faut noter ici que l'importance donnée aux nouveaux gènes dans la diversification fonctionnelle des génomes ne doit pas faire oublier un second mécanisme majeur qui participe également aux innovations fonctionnelles : l'épissage alternatif des gènes (Kopelman *et al.* 2005). Néanmoins, le présent manuscrit se concentre sur le rôle des duplications de gènes et de génomes dans l'évolution phénotypique des organismes vivants, et en particulier chez les vertébrés.

En 1970, le généticien Susumu Ohno attribue, dans son célèbre livre *Evolution by Gene Duplication*, un rôle fondamental aux duplications de gènes pour l'innovation génétique. Comme mentionnée ci-dessus, cette idée, proposée à l'ère pré-moléculaire, est l'aboutissement d'une longue réflexion initiée dès le début du XXe siècle (Haldane 1933, Muller 1935, Ohno 1970, pour revue voir Taylor & Raes 2004). De ce point de vue, une seconde copie d'un gène constitue une opportunité pour diversifier le génome du vivant. En effet, une des deux copies conserve la fonction originale, tandis que la seconde évolue plus librement, avec accumulation de mutations, et l'éventualité d'acquérir de nouvelles propriétés fonctionnelles. Depuis la publication du livre de Ohno, et suite à l'avènement de l'ère moléculaire, cette hypothèse pionnière est confortée par de nombreux travaux

(Conant & Wolfe 2008). Par ailleurs, le généticien propose que l'augmentation de la taille du génome chez les vertébrés soit la conséquence de deux duplications de génome complet survenues chez l'ancêtre commun des vertébrés (Ohno 1970). Ohno établit alors un lien de cause à effet entre la complexification du génome et les innovations évolutives survenues chez les vertébrés. L'avènement de l'ère génomique et les études engagées ces dernières années confirment l'occurrence de duplications de génome au cours de l'évolution des vertébrés, et plus généralement des eucaryotes (Van de Peer *et al.* 2009). Finalement, la communauté scientifique propose un rôle majeur aux duplications de gènes et de génomes dans l'évolution moléculaire et morphologique des animaux (Cañestro *et al.* 2013). Cependant, ces questions sont encore à l'étude et en grande partie non résolues. Ainsi, de très nombreuses études sont menées pour comprendre les mécanismes moléculaires qui provoquent et/ou facilitent la duplication d'un gène, la préservation des gènes en deux copies, et l'acquisition de nouvelles fonctions, et donc pour améliorer notre compréhension des bases génétiques de la diversification des organismes vivants.

Comment dupliquer un gène ?

La duplication est le processus par lequel un gène, un fragment de génome ou un génome entier passe d'une à deux copies, initialement identiques, mais qui peuvent, après fixation dans la population, diverger au cours de l'évolution. L'existence de duplications nombreuses et de longueurs variées est très visible dans les génomes. En particulier, les génomes eucaryotes contiennent de nombreuses familles de gènes homologues (familles multigéniques) produites par duplication puis divergence à l'intérieur des génomes (Orengo & Thornton 2005). Les gènes homologues produits par un événement de duplication sont appelés «paralogues» pour les distinguer des «orthologues» issus d'un événement de spéciation. Les types de duplication peuvent être classés en deux catégories: les duplications à petite échelle, ou duplications segmentaires, et les duplications à grande échelle, ou duplications de génome. Chez les eucaryotes, les duplications segmentaires se produisent en continu, et par des mécanismes moléculaires divers. Ainsi, les duplications de gènes semblent être aussi fréquentes que la substitution d'un nucléotide par un autre (Lynch & Conery 2000). Plus rarement, la totalité des gènes présents dans un génome peut être dupliquée au cours d'une duplication de génome complet. En effet, de nombreux eucaryotes présentent des traces de duplications génomiques ancestrales, notamment chez les levures, les plantes et les vertébrés (Van de Peer *et al.* 2009). Ce chapitre est destiné à présenter brièvement les différents types de duplication segmentaire, et introduire le chapitre suivant qui traitera des duplications de génome (polyploïdisation).

Duplication de gène par duplication segmentaire

Les duplications segmentaires sont produites par des mécanismes moléculaires variés (Lynch 2007). Les duplications segmentaires médiées par un intermédiaire d'acide désoxyribonucléique (ADN), considérées jusqu'à récemment comme majoritaires, sont les plus étudiées (figure 6). Ce type de duplication est essentiellement la conséquence de recombinaisons non-homologues survenues au cours de la méiose, et provoque habituellement des duplications de gènes dites en tandem, avec deux copies dupliquées côte-à-côte. Par exemple, le complexe des gènes *Hox* retrouvés chez les animaux invertébrés, et impliqué dans le développement embryonnaire, est le résultat de plusieurs événements de duplication en tandem survenus chez l'ancêtre commun des

bilatériens. Des réarrangements chromosomiques subséquents peuvent néanmoins provoquer la dispersion dans le génome de ces duplications initialement locales. Plus récemment, de nombreuses études montrent l'importance des duplications médiées par un intermédiaire acide ribonucléique (ARN) pour la production de nouveaux gènes (Kaessmann 2010). Ce type de duplication est nommé rétroposition ou rétroduplication. Dans ce cas, un ARN messenger mature (ARNm), transcrit à partir d'un gène «source», est rétrotranscrit en ADN complémentaire puis inséré dans le génome. Ce mécanisme est permis par la présence dans le cytoplasme des enzymes de rétroposition du fait de l'activité de certains rétrotransposons. Le nouveau gène, issu d'un ARNm mature (dont les introns sont épissés), ne contient que les exons du gène «source». Ce mécanisme, qui crée une copie dépourvue d'intron et de séquences régulatrices, produit fréquemment des pseudogènes (gènes inactifs dans un génome). Néanmoins, l'avènement de l'ère génomique a permis d'identifier de nombreuses rétrocopies fonctionnelles (nommées rétrogènes) dans les génomes. Ces rétrogènes ont souvent un rôle dans les cellules germinales, mais également dans d'autres processus, tels que le comportement de cour chez la drosophile, ou la longueur des jambes chez le chien (Kaessmann 2009). La petite taille des jambes de certains chiens (le teckel par exemple) est en effet provoquée par la présence d'une rétrocopie du gène *fgf4*, dotée d'une queue polyA et dépourvue d'introns, et exprimée au cours du développement de l'animal (Parker *et al.* 2009).

Duplication de gène par duplication de génome

Le paragraphe précédent présente des mécanismes moléculaires qui dupliquent un unique gène, ou quelques gènes, dans un génome. Cependant, un cas extrême de mutation peut provoquer la duplication de la totalité des gènes présents dans un génome : un événement de polyploïdisation, ou la duplication complète d'un génome. Cet événement constitue une façon simple et rapide de dupliquer de très nombreux gènes, mais également de permettre la duplication de voies métaboliques complètes (Conant & Wolfe 2008). De nombreuses espèces actuelles sont polyploïdes, notamment des plantes, des poissons téléostéens et des amphibiens. Ainsi, l'évolution des eucaryotes est marquée par plusieurs événements de duplication de génome complet, particulièrement chez les plantes, mais également chez les animaux. Le chapitre suivant, qui traite de la polyploïdie, se décompose en quatre parties. La première partie présente les mécanismes moléculaires qui provoquent la duplication d'un génome complet, et le devenir d'une telle mutation dans une population diploïde. La seconde partie expose les

hypothèses actuelles qui expliquent le succès évolutif de certaines populations polyploïdes ancestrales, notamment chez les plantes et les vertébrés. Enfin, les deux dernières parties rappellent les méthodes pour identifier ces polyploïdes ancestraux (actuellement redevenus diploïdes) et présentent les différents événements de polyploïdisation identifiés au cours de l'évolution des vertébrés.

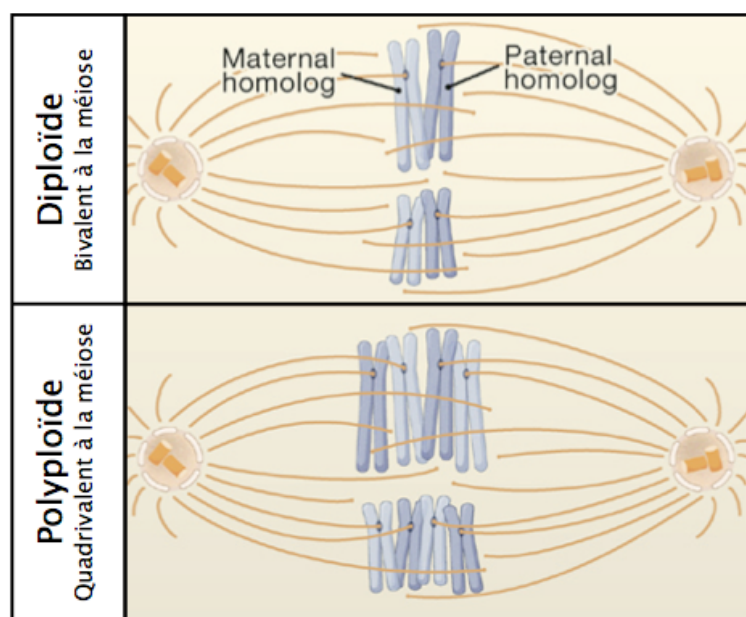


Figure 7: Génomes diploïde et polyploïde (Otto 2007). Dans le génome diploïde, deux chromosomes non-homologues sont représentés en forme de croix (deux chromatides sœurs jointes au niveau du centromère). Dans le génome tétraploïde, juste après duplication de génome, les chromosomes forment un quadrivalent à la méiose.

Polyploïdie et duplications de génome

Un cas extrême de mutation : la polyploïdisation ou la duplication complète d'un génome

La polyploïdisation est un événement de mutation qui engendre le passage d'un état diploïde ($2n$ chromosomes) à un état polyploïde (plus de $2n$ chromosomes), et par conséquent la duplication complète du génome (figure 7). La polyploïdie peut être obtenue par autopolyploïdisation (événement intra-spécifique) ou par allopolyploïdisation (hybridation inter-spécifique). Dans les deux cas, une mauvaise disjonction des chromosomes homologues au cours de la méiose aboutit à un défaut de réduction gamétique, et à l'établissement d'un individu polyploïde. Chez les plantes, les polyploïdes sont produits à une fréquence relativement élevée de 1 pour 100 000 individus (Ramsey & Schemske 1998). La polyploïdie ne perturbe pas la mitose, et un zygote polyploïde peut se développer sans anomalie. En revanche, la polyploïdie peut avoir des conséquences sur la méiose, et en cela sur la fertilité des individus polyploïdes, lorsque le nombre de chaque type de chromosome est impair (individus triploïdes ($3n$) par exemple). La tétraploïdie ($4n$) quant à elle n'affecte pas le déroulement de la méiose et les individus sont fertiles. Néanmoins, un individu tétraploïde doit se reproduire avec d'autres individus tétraploïdes pour obtenir une descendance capable de réaliser la méiose, et ainsi la production de gamètes viables. Finalement, dans une population d'individus diploïdes, quelques rares tétraploïdes peuvent se croiser entre eux ou s'autoféconder, et ainsi former une population présentant, de part leur constitution génétique, un isolement reproductif *quasi* total. La population polyploïde peut alors supplanter la population ancestrale lorsque sa particularité génétique est favorisée dans un environnement donné. En effet, à court terme, les polyploïdes peuvent présenter, dans un écosystème perturbé, un avantage compétitif face à leurs parents diploïdes, la duplication du génome favorisant notamment la redondance fonctionnelle et la robustesse mutationnelle (Crow & Wagner 2006). En outre, l'effet d'hétérosis, et l'augmentation de la variabilité phénotypique, peut permettre aux polyploïdes d'envahir de nouvelles niches écologiques, jusqu'alors inaccessibles aux parents diploïdes (Hegarty & Hiscock 2007). C'est ainsi que de très nombreuses espèces de plantes, et la plupart des plantes cultivées, sont polyploïdes. Chez les animaux, bien que plus rare, des

centaines d'espèces d'insectes et de vertébrés (principalement des amphibiens et des poissons) sont également polyploïdes (Van de Peer *et al.* 2009).

Polyploïdie : une impasse évolutive ?

Bien que la production d'individus polyploïdes chez les eucaryotes soit relativement fréquente, et à l'origine de nombreuses espèces polyploïdes actuelles, les événements de polyploïdisation ancestraux (paléopolyploïdisation) semblent extrêmement rares. Très peu de populations ancestrales polyploïdes semblent s'être durablement établies au cours de l'évolution des eucaryotes. Par exemple, les plantes à fleurs semblent partager uniquement quatre duplications de génome complet depuis leur émergence il y a 150 à 200 millions d'années (Tang *et al.* 2008, Jaillon *et al.* 2007). Les vertébrés partagent quand à eux seulement deux duplications de génome complet (et une troisième additionnelle chez les téléostéens) survenues durant les derniers 500 à 600 millions d'années (voir ci-dessous). Ce constat suggère que la polyploïdie est généralement une impasse évolutive, mais que dans certaines conditions environnementales, elle peut aboutir à la préservation et au succès évolutif des populations polyploïdes sur le long terme. En effet, Van de Peer et ses collègues suggèrent que des bouleversements drastiques de l'environnement, favorisant l'apparition de nouvelles niches écologiques et perturbant les écosystèmes établis, sont des événements contingents déterminants pour la survie et le succès évolutif à long terme des duplications de génome complet (Van de Peer *et al.* 2009). Ainsi, pour appuyer leur hypothèse, ils rappellent que les deux duplications de génome survenues chez l'ancêtre commun des vertébrés (520-550 Ma, voir plus bas) sont proches de l'extinction de masse du Cambrien (542 Ma). De même, la duplication de génome spécifique aux téléostéens datant du dévonien (226-316 Ma) est proche de l'extinction Permien-Trias (250 Ma). Enfin, les nombreuses duplications de génome indépendantes préservées chez différentes espèces de plantes (légumineuses, céréales et solanacées) sont survenues proches de l'extinction Crétacé-Tertiaire il y a 65 Ma. Ainsi, Van de Peer et ses collègues proposent un scénario adaptatif pour expliquer le succès évolutif à court terme des polyploïdes actuels (développé dans le précédent paragraphe), et attribuent un rôle fondamental aux perturbations drastiques de l'environnement pour expliquer l'établissement et le succès évolutif à long terme des paléopolyploïdes.

Comment identifier les génomes paléopolyploïdes ?

L'identification des événements de polyploïdisation est un défi pour les chercheurs, particulièrement lorsque l'événement est ancien. Les duplications de génome récentes peuvent être détectées en identifiant les espèces dont le caryotype contient deux fois plus de chromosomes que leurs espèces voisines. En outre, les événements de polyploïdisation récents sont identifiables facilement par la présence de grandes régions dupliquées dans les génomes des espèces concernées. En revanche, pour les événements anciens, il est plus difficile d'identifier de telles régions, dans la mesure où de nombreux réarrangements génomiques (délétions de gènes, événements de translocation et recombinaisons) sont survenues. Ainsi, les génomes paléopolyploïdes sont certes structurellement polyploïdes, mais redevenus génétiquement diploïdes suite à de nombreux réarrangement génomiques, et ainsi difficile à identifier. Cependant, des outils informatiques ont été développés ces dernières années pour détecter des régions anciennement dupliquées dans les génomes qui à première vue ne semblent pourtant présenter aucune homologie (Van de Peer 2004). En particulier, ces outils permettent l'identification dans les génomes de larges régions chromosomiques contenant des ensembles de gènes homologues (et souvent produits par duplication de génome). Ainsi, lorsque la similarité entre deux régions chromosomiques est significative (c'est à dire que la vraisemblance qu'elles soient le résultat du hasard est *quasi* nulle), les gènes dupliqués qui composent ces deux régions chromosomiques sont considérés comme la conséquence d'un unique événement de duplication. En pratique, pour détecter des régions homologues dans les génomes, les chromosomes sont représentés par des listes de gènes ordonnés selon leur position sur ces chromosomes. A partir de ces listes de gènes, les gènes qui présentent des homologues dans d'autres régions chromosomiques sont identifiés par BLAST (Basic Local Alignment Search Tool ; Altschul *et al.* 1990). Il faut noter ici que les très grandes familles de gènes peuvent rendre difficile l'identification des régions du génome véritablement homologues (bruit de fond). Le retrait de certaines grandes familles (après avoir défini une taille seuil) avant analyse, ou le calcul de l'espérance mathématique que deux gènes soient paralogues sont deux solutions généralement envisagées. Finalement, les informations recueillies sur les gènes homologues sont alors rassemblées au sein d'une matrice de gènes homologues. Dans une telle matrice, les régions chromosomiques colinéaires (même contenu et même ordre de gènes) apparaissaient sous la forme de diagonale (figure 8). C'est grâce à cette méthode que Wolfe et Shields identifient en 1997 de nombreux blocs de gènes dupliqués

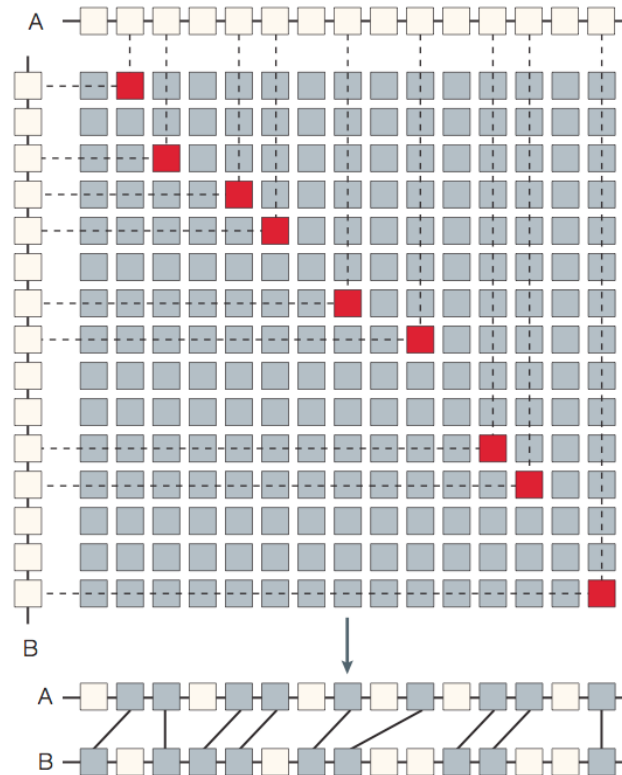


Figure 8: Matrice de gènes homologues (Van de Peer 2004). A et B représentent deux régions chromosomiques hypothétiques. Chaque carré représente un gène. Les carrés rouges dans la matrice de gènes homologues indiquent les gènes homologues entre les deux régions A et B. Ces derniers sont représentés par des carrés gris reliés par des traits noirs dans le résultat d'analyse au dessous. Lorsque les carrés rouges forment une diagonale, les gènes homologues correspondant font partie d'une région chromosomique dupliquée. Lorsque les carrés rouges forment une horizontale ou une verticale (non montré dans le schéma), les gènes homologues sont le produit de duplications en tandem sur l'une ou l'autre des deux régions.

chez la levure *Saccharomyces cerevisiae* (Wolfe & Shields 1997). Cette étude montre ainsi que le génome de la levure est génétiquement diploïde, mais structurellement polyplœide, et en cela paléopolyplœide. Depuis, des génomes paléopolyplœides ont été identifiés chez de nombreux autres eucaryotes, notamment chez le cilié *Paramecium tetraurelia*, qui a subi plusieurs événements de polyplœidisation, mais également chez les plantes et les animaux vertébrés (Van de Peer *et al.* 2009).

Paléopolyplœidie chez les vertébrés

C'est au cours des années 1970 qu'Ohno formule une hypothèse pionnière pour expliquer l'augmentation de la taille des génomes chez les vertébrés: un, deux, ou même trois événements de polyplœidisation sont survenus au cours de l'évolution des vertébrés (Ohno 1970). Au cours des années 1990, cette hypothèse est reprise et gagne en popularité, en particulier parmi les biologistes du développement. Cependant, les supports empiriques pour soutenir cette hypothèse restent faibles et reposent essentiellement sur l'analyse de quelques familles de gènes, principalement les gènes *Hox*. Ainsi, les protostomes et les deutérostomes invertébrés possèdent généralement un unique gène lorsque les génomes de vertébrés en contiennent typiquement plusieurs, souvent quatre copies de ce gène (Holland *et al.* 1994, Meyer 1999). En outre, le nombre et le moment des événements de duplication au cours de l'évolution des vertébrés varient selon les auteurs (Lundin 1993, Holland *et al.* 1994, Kasahara *et al.* 1996, Meyer & Schartl 1999). Cependant, au début des années 2000, l'hypothèse la plus populaire place une première duplication de génome à la base des animaux vertébrés, et une seconde à la base des gnathostomes (Makałowski 2001). Finalement, au cours des années 2000, le séquençage de plusieurs génomes de chordés, le développement des outils informatiques (dont certains présentés dans le paragraphe précédent) pour l'étude des génomes, la découverte subséquente de nombreuses régions chromosomiques paralogues, et plus récemment la comparaison des séquences complètes des génomes de différents chordés, ensemble supportent fortement l'hypothèse des deux duplications de génome survenues au cours de l'évolution des chordés dans la lignée des vertébrés (Hoegg & Meyer 2005, Dehal & Boore 2005, Putnam *et al.* 2008, Van de Peer *et al.* 2010). Ainsi, actuellement, la majorité des chercheurs (mais pas la totalité, voir Abbasi 2010) acceptent l'hypothèse des deux duplications de génomes à la base des vertébrés (Van de Peer *et al.* 2010). En revanche,

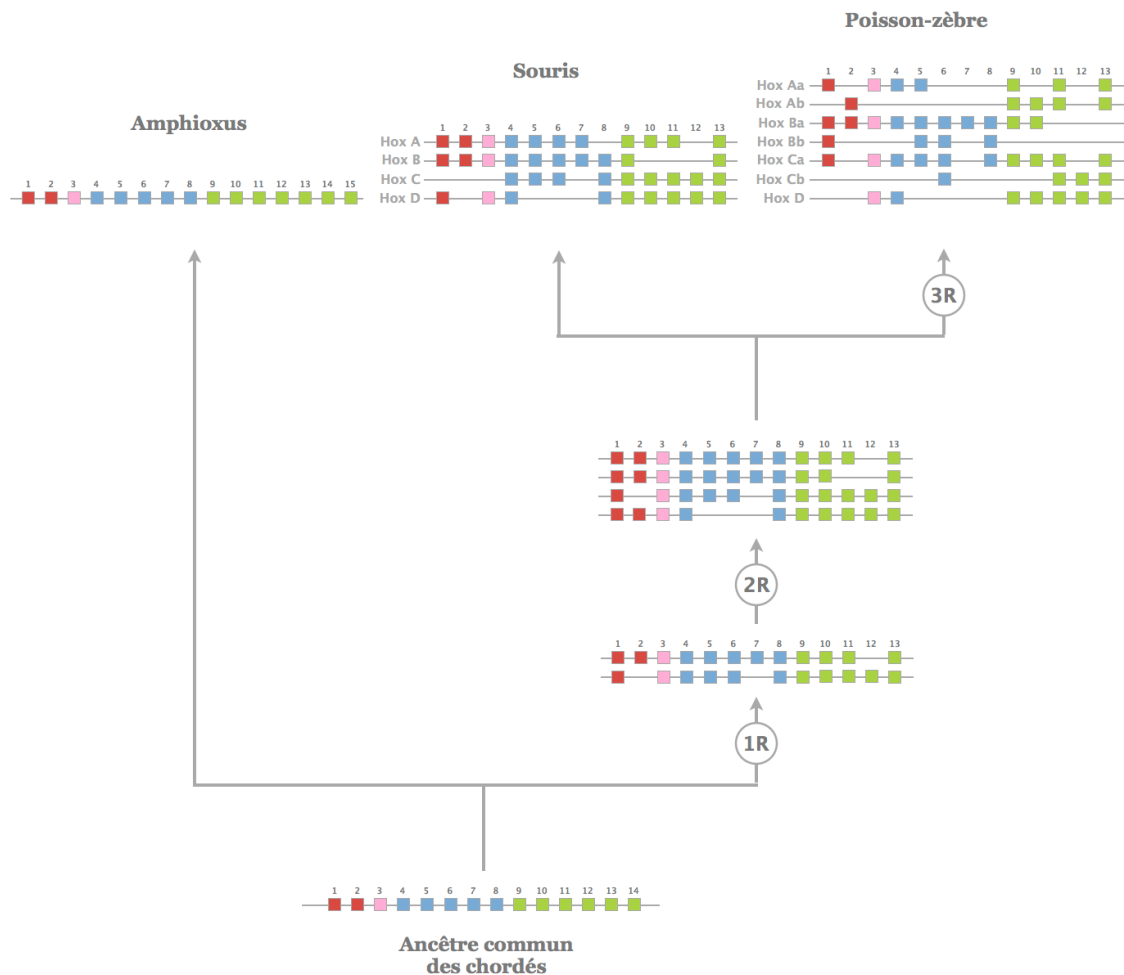


Figure 9: Scénario évolutif des complexes *Hox* chez les chordés. Les différents événements de polypléidisation sont indiqués par 1R, 2R et 3R (d'après Prince & Pickett 2002).

les conséquences évolutives des duplications de génome, notamment leurs rôles dans la diversification des phénotypes et les innovations évolutives qui caractérisent les vertébrés restent des questions ouvertes et en grande partie non résolues (Crow & Wagner 2006). Enfin, la datation des deux événements (520-550 Ma) est encore un sujet de débat : tandis que la communauté scientifique s'accorde à placer la première duplication de génome à la base des vertébrés, la seconde duplication de génome est positionnée soit à la base des vertébrés, soit à la base des gnathostomes, après la divergence avec les agnathes (Kuraku *et al.* 2009). Cependant, le séquençage récent du génome de la lamproie suggère que les deux événements de duplication sont survenus à la base des vertébrés (Smith *et al.* 2013).

A la fin des années 1990, la découverte des complexes *Hox* surnuméraires chez le poisson-zèbre, puis chez le fugu et le médaka, conduit les scientifiques à inférer un événement de polyploïdisation additionnel à la base des actinoptérygiens (Amores *et al.* 1998, Meyer & Schartl 1999, Aparicio 2000, Naruse *et al.* 2000). En effet, ces travaux montrent que les sarcoptérygiens possèdent quatre complexes *Hox* sur quatre chromosomes différents, tandis que les génomes des poisson-zèbre et médaka contiennent sept complexes *Hox* également présents sur différents chromosomes (figure 9, Amores *et al.* 1998, Naruse *et al.* 2000). L'étude subséquente de l'histoire évolutive de nombreux gènes chez les vertébrés semble confirmer cette hypothèse (Taylor *et al.* 2001). Finalement, le séquençage du génome du tétraodon en 2004, et une approche de génomique comparative avec le génome humain, ont permis de confirmer cette hypothèse, et d'inférer l'événement de polyploïdisation à la base des téléostéens (Jaillon *et al.* 2004). Actuellement, la duplication de génome spécifique aux téléostéens est datée entre 226 et 350 Ma (Van de Peer *et al.* 2009).

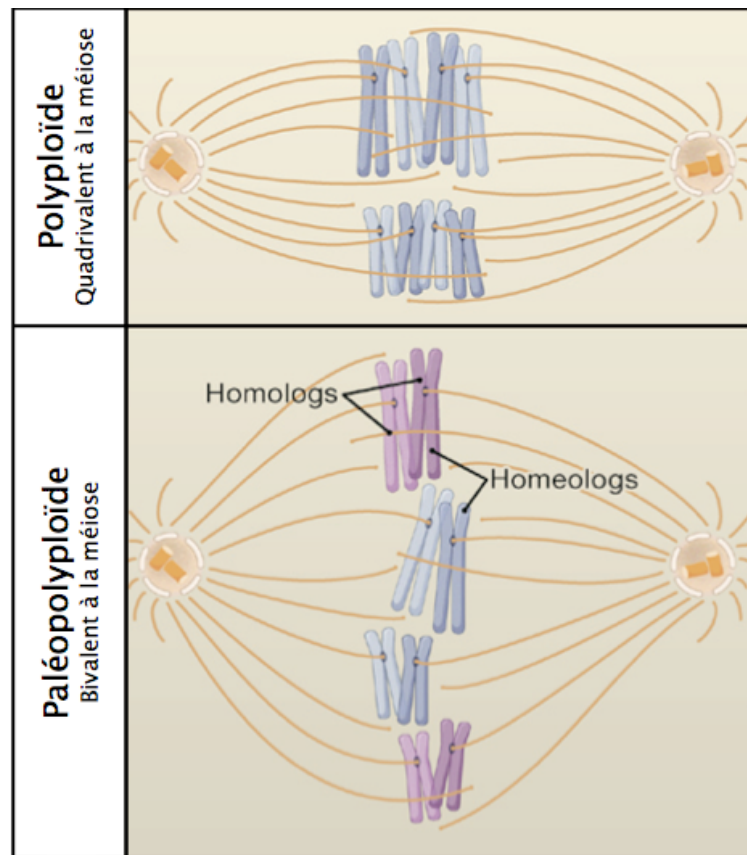


Figure 10: Diploïdisation (Otto 2007). Au cours de la méiose, la diploïdisation se traduit par le passage d'un quadrivalent, avec quatre chromosomes homologues à l'équateur de la cellule, à un bivalent formé de deux chromosomes homologues. Les chromosomes homéologues sont deux chromosomes paralogues qui dérivent d'un même proto-chromosome.

Quel devenir pour les génomes et les gènes dupliqués ?

Le mystère de la diploïdisation

Les génomes anciennement polyploïdes sont actuellement diploïdes. Ce constat révèle un aspect de l'évolution des génomes encore mal compris: le processus de diploïdisation. En effet, les espèces polyploïdes, qui évitent l'extinction, adoptent sur le long terme une trajectoire évolutive nommée diploïdisation, qui consiste en la conversion graduelle de la polyploïdie à la diploïdie. Au cours de la méiose, la diploïdisation se traduit par le passage d'un quadrivalent, formé de quatre chromosomes homologues à l'équateur de la cellule, à un bivalent, composé de deux chromosomes homologues (figure 10). Les bases moléculaires d'un tel processus sont peu comprises, mais impliquent de nombreux réarrangements génomiques, et semblent associées à des événements de délétion (dont de nombreux gènes) et une réduction de la redondance génomique (Wolfe 2001, Comai 2005). Sémon et Wolfe ont étudié l'évolution du taux de réarrangement génomique après la duplication de génome complet survenue chez l'ancêtre commun des téléostéens (Sémon & Wolfe 2007). Ils concluent que ce taux tend à augmenter après la duplication de génome, en accord avec le processus de diploïdisation. Par la suite, la reconstruction du contenu chromosomique de l'ancêtre commun des téléostéens à partir des génomes des trois médaka, tétraodon et poisson-zèbre a permis d'inférer l'histoire des réarrangements génomiques survenus dans ces différentes lignées suite à la duplication de génome (Kasahara *et al.* 2007). Les auteurs montrent que huit réarrangements génomiques majeurs se sont déroulés chez l'ancêtre commun des téléostéens juste après la duplication de génome. Ces données confirment une augmentation transitoire du taux de réarrangements génomiques juste après duplication de génome, au cours de la diploïdisation. Cette étude indique également la présence de réarrangements inter-chromosomiques lignée-spécifiques, ce qui suggère des événements de délétion et de réorganisation génomique particuliers pour chaque espèce. En conséquence, les changements génomiques qui marquent le passage de la polyploïdie à la diploïdie, ou diploïdisation, participent grandement à la modulation contingente du contenu génomique, quelque fois de façon lignée-spécifique, et notamment au devenir des gènes dupliqués par duplication de génome.

Un gène ou deux ?

Initialement, un événement de polyploïdisation double le contenu du génome, et par conséquent le contenu génique. Cependant, après un événement de polyploïdisation, les génomes polyploïdes subissent de nombreux événements de délétion au cours du processus de diploïdisation présenté précédemment, ce qui réduit le contenu génique. De plus, les deux copies d'un gène produites par duplication de génome sont redondantes, et, généralement, une des deux copies, qui subit un relâchement de la sélection négative, devient non-fonctionnelle par accumulation de nombreuses mutations (Lynch & Conery 2000). Ce processus, nommé pseudogénisation, est multifactoriel, et implique des mutations au niveau de la séquence codante du gène, et/ou des éléments régulateurs qui gouvernent l'expression de ce gène, mais également des modifications épigénétiques (Chain & Evans 2006). Ainsi, après un événement de duplication de génome, la majorité des nouvelles copies tend à être perdue du génome par délétion ou pseudogénisation (figure 12a). En effet, les travaux réalisés dans divers taxons (levures, plantes, vertébrés) révèlent que seulement 10% à 30% des gènes sont retenus en deux copies après une duplication de génome (Kellis *et al.* 2004, Bowers *et al.* 2003, Brunet *et al.* 2006). De plus, la majorité des événements de non-fonctionnalisation semble survenir rapidement après duplication (certainement au cours du processus de diploïdisation), ce qui explique des pourcentages de perte similaires quelque soit l'ancienneté de l'événement de duplication (Brunet *et al.* 2006). Néanmoins, ces observations suggèrent l'existence de mécanismes de préservation qui contrebalancent ou exploitent l'accumulation de mutations dans les séquences codantes et/ou les éléments régulateurs, et protègent les deux copies d'un événement de délétion. D'une part, la nécessité de préserver les quantités stœchiométriques, ou encore un équilibre du dosage, entre les gènes impliqués dans une même unité fonctionnelle après duplication de génome favoriserait la rétention en deux copies de ces gènes sensibles à l'effet de dosage (Freeling & Thomas 2006, Makino & McLysaght 2010). Cette hypothèse, initialement formulée par Papp *et al.* (2003), permet d'expliquer notamment la sur-représentation de certaines catégories fonctionnelles après duplication de génome (voir ci-après, Freeling & Thomas 2006). D'autre part, les modèles classiques de sous-fonctionnalisation et néo-fonctionnalisation, présentés ci-dessous, permettent également d'expliquer la préservation de nombreux gènes dupliqués dans les génomes eucaryotes (Conant & Wolfe 2008). Finalement, après préservation, les nouvelles propriétés fonctionnelles d'une copie semblent très souvent être des exaptations de

fonctions pré-existantes, co-optées plus ou moins longtemps après l'événement de duplication (Conant & Wolfe 2008, idée développée dans la dernière partie de ce chapitre).

Les mécanismes de préservation des gènes dupliqués

La préservation des gènes dupliqués par effet de dosage

Cette hypothèse se base sur le postulat que la quantité de transcrits produite par un gène (et de produit protéique correspondant) est directement proportionnelle au nombre de copies de ce gène présentes dans le génome. La relation entre le nombre de copies d'un gène et la quantité de produit protéique s'appelle l'effet de dosage des gènes. L'hypothèse postule un équilibre du dosage relatif des gènes dont les produits interagissent ou sont impliqués dans une même unité fonctionnelle (définie ici comme un complexe protéique, une voie métabolique, une cascade de signalisation, ou un réseau de régulation ; Freeling & Thomas 2006). Ainsi, sous cette hypothèse, si le dosage relatif de certains gènes change, par exemple suite à un événement de duplication (ou de délétion) segmentaire, des déséquilibres physiologiques peuvent apparaître dans l'unité fonctionnelle correspondante et en cela être délétères (figure 11, Makino & McLysaght 2010). Par conséquent, les gènes sensibles à l'effet de dosage sont généralement des gènes «résistants» aux duplications segmentaires, car l'effet délétère immédiat de la duplication segmentaire de ce type de gène empêche la fixation de la nouvelle copie dans la population (Conant & Wolfe 2008). En revanche, un événement de polyploïdisation fournit une unique opportunité pour dupliquer les gènes sensibles à l'effet de dosage (et la fixation subséquente des deux copies dans la population polyploïde) car elle garantit la duplication simultanée de l'ensemble des gènes impliqués dans une même unité fonctionnelle et le maintien de l'équilibre du dosage (Makino & McLysaght 2010). En outre, comme mentionné dans le titre de ce paragraphe, l'effet de dosage favorise la rétention post-fixation en deux copies de ces gènes après duplication de génome. En effet, de la même manière qu'un événement de duplication segmentaire peut changer le dosage relatif des gènes impliqués dans une même unité fonctionnelle, la délétion d'un gène particulier, après duplication de l'ensemble de l'unité, peut produire également un déséquilibre du dosage. Ainsi, le maintien de l'équilibre du dosage (par sélection négative) conduit mécaniquement à la préservation en deux copies des gènes sensibles à l'effet de dosage sur de longues périodes évolutives

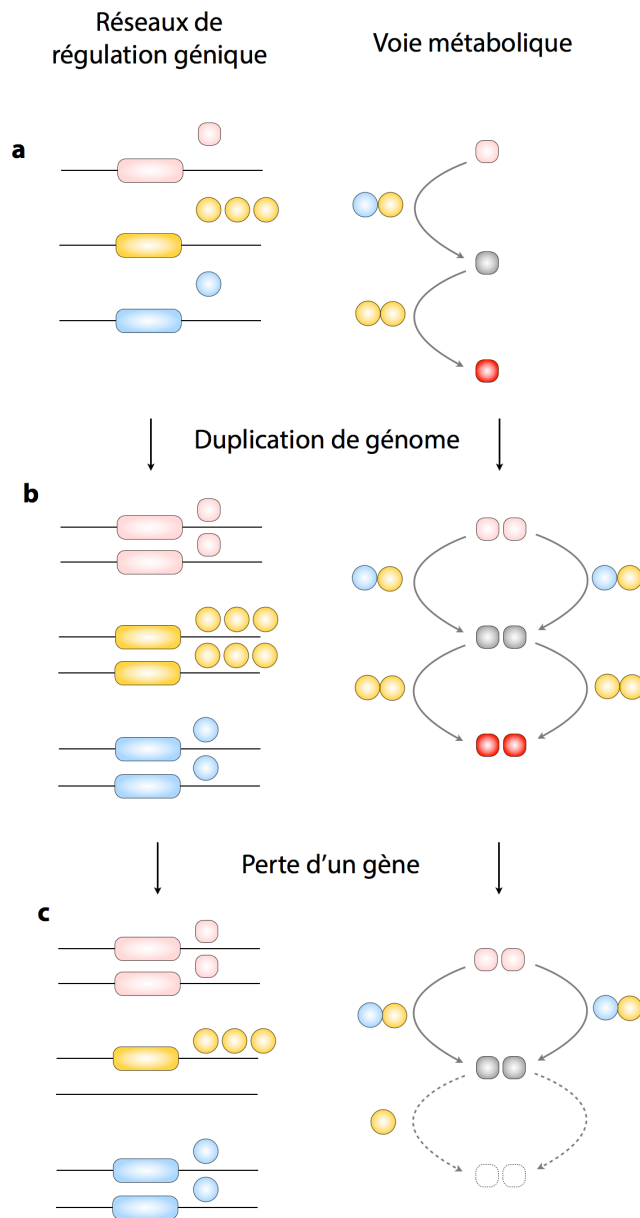


Figure 11: Duplication de génome et effet de dosage. Les produits de trois gènes interagissent en quantité stœchiométrique dans une même voie métabolique (a). Après duplication de génome, l'équilibre du dosage des trois gènes est maintenu (b). La perte d'un gène impliqué dans la voie métabolique altère l'équilibre du dosage et peut entraîner des déséquilibres physiologiques (c).

(Freeling & Thomas 2006). En accord avec cette hypothèse, de nombreux travaux montrent que, dans les génomes eucaryotes, les gènes dupliqués par duplication de génome appartiennent majoritairement à certaines catégories fonctionnelles, en particulier les gènes sensibles à l'effet de dosage: les gènes codant les protéines de signalisation cellulaire et les facteurs de transcription, et plus généralement les gènes du développement (Taylor & Raes 2004, Blanc & Wolfe 2004, Maere *et al.* 2005, Wapinski *et al.* 2007). Enfin, l'hypothèse développée dans ce paragraphe implique l'idée que les événements de polyploïdisation peuvent conduire à certaines innovations évolutives, notamment de part l'enrichissement de certaines catégories fonctionnelles (sensibles au dosage), qui auraient eu peu de chance de se produire autrement (idée abordée plus en détails dans le prochain chapitre).

La préservation par néo-fonctionnalisation

Ce modèle, proposé par Ohno en 1970, postule que, après duplication d'un gène, une des deux copies conserve la fonction ancestrale, tandis que la seconde copie, qui subit un relâchement de la sélection négative (car redondante), évolue plus librement et fixe de plus nombreuses mutations par dérive génétique. Comme mentionné plus haut, l'accumulation de mutations, souvent défavorables, aboutit très fréquemment à la progressive pseudogénisation de la seconde copie, et finalement à sa perte. Néanmoins, certaines rares mutations avantageuses peuvent être retenues sous l'effet de la sélection naturelle (sélection positive), et conférer de nouvelles propriétés fonctionnelles à la seconde copie (figure 12b). La néo-fonctionnalisation, qui conduit à la préservation sur le long terme des deux copies dans le génome, est un modèle à présent supporté par de nombreuses données expérimentales (Conant & Wolfe 2008). Chez le cochon, la duplication du gène cytochrome P450arom illustre le modèle de néo-fonctionnalisation. En effet, suite à l'événement de duplication, un paralogue réalise la fonction originale (synthèse d'œstrogène à partir d'androgène) tandis que la seconde copie effectue une nouvelle fonction (activité hydroxylase) au niveau des gonades de l'animal (Corbin *et al.* 2004). Cette nouvelle fonction (ou néo-fonctionnalisation), absente chez les autres mammifères, permet la synthèse de la 1-beta-hydroxytestostérone, impliquée dans la biologie de la reproduction du cochon. Les méthodes basées sur les séquences des gènes sont très utilisées pour inférer des événements de néo-fonctionnalisation. En effet, sous cette hypothèse, les mutations à l'origine des nouvelles propriétés fonctionnelles se sont fixées (par sélection positive) plus rapidement dans la population que les mutations neutres. Ainsi, le taux de substitutions non-synonymes (dN) est en théorie

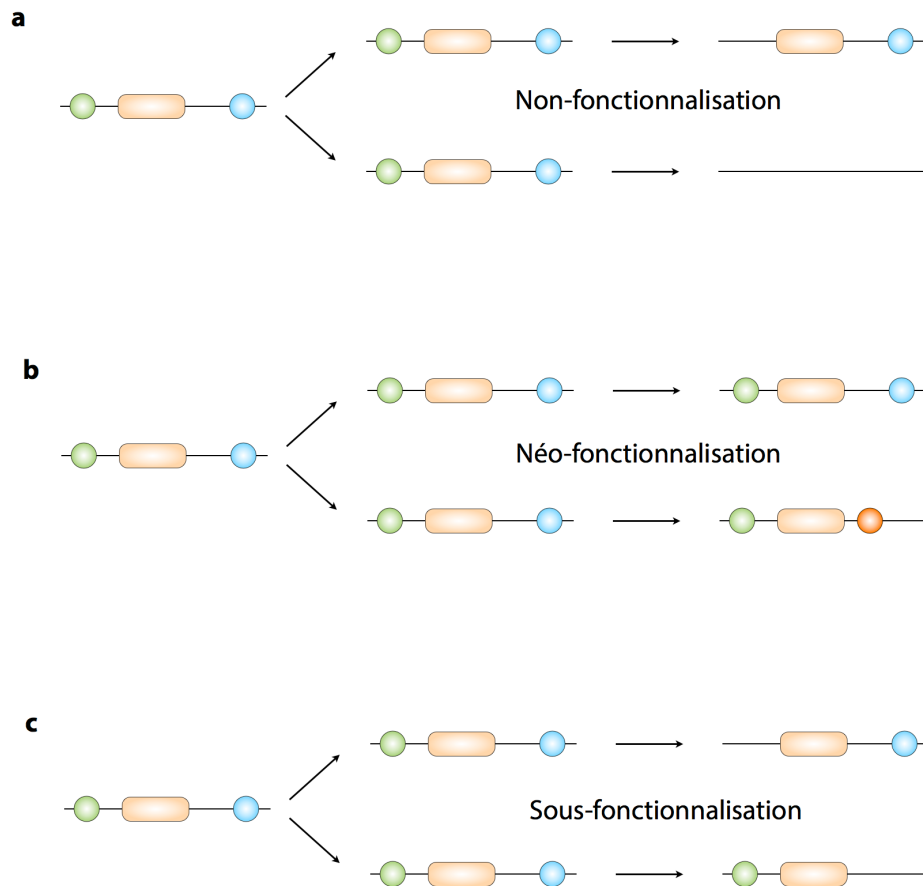


Figure 12: Mécanismes de préservation des gènes dupliqués. Les cylindres marrons clairs représentent la séquence codante d'un gène et les ronds verts, bleus et rouges représentent les éléments *cis*-régulateurs qui gouvernent l'expression de ces gènes. La non-fonctionnalisation conduit à la perte d'une des deux copies après une période de redondance (a). La néo-fonctionnalisation, ici représentée par l'acquisition d'un élément *cis*-régulateur, correspond à une divergence fonctionnelle d'une des deux copies (b). La sous-fonctionnalisation correspond au partage des sous-fonctions ancestrales (ici des domaines d'expression) entre les deux gènes dupliqués (c). Les modèles de néo-fonctionnalisation et la sous-fonctionnalisation peuvent inclure également des changements au niveau de la protéine.

significativement supérieur au taux de substitutions synonymes (dS). Par conséquent, les patrons de variations non-synonyme et synonyme peuvent permettre de caractériser les effets de la sélection naturelle, et ainsi l'identification de cas de néo-fonctionnalisation après un événement de duplication. Néanmoins, la réciproque n'est pas vraie, et certains cas de néo-fonctionnalisation seront difficilement identifiables avec les seules données de séquences, les signatures moléculaires de sélection positive ayant pu être effacées au cours du temps (Kreitman & Akashi 1995, Chain & Evans 2006). De plus, une évolution adaptative est compatible avec d'autres modèles d'évolution des gènes dupliqués, notamment dans certains cas de sous-fonctionnalisation (comme mentionné ci-dessous), et conclure à un scénario de néo-fonctionnalisation peut quelquefois s'avérer très spéculatif. Enfin, la probabilité d'occurrence d'une mutation avantageuse avant qu'une mutation délétère n'aboutisse à la perte de l'un des paralogues est faible (et donc de préservation par néo-fonctionnalisation). Pourtant, de très nombreux gènes sont retenus en deux ou plusieurs copies dans les génomes, notamment ceux appartenant aux familles multigéniques, ce qui est difficilement explicable par le seul modèle de néo-fonctionnalisation. Le modèle de sous-fonctionnalisation, proposé par différents auteurs au cours des années 1990, permet d'expliquer plus amplement la présence de très nombreux gènes dupliqués dans les génomes eucaryotes (Piatigorsky *et al.* 1988, Hughes 1994, Force *et al.* 1999, Taylor & Raes 2004).

La préservation par sous-fonctionnalisation

Les deux paralogues *SIR3* et *ORC1* présents dans le génome de la levure *Saccharomyces cerevisiae* sont les produits d'une duplication de génome complet. Ces deux gènes possèdent deux fonctions différentes chez cette espèce de levure. Les travaux de Van Hoof montrent que la protéine orthologue présente chez la levure *Saccharomyces kluyveri* (qui n'a pas subi de duplication de génome) est capable de réaliser la fonction des deux copies *SIR3* et *ORC1* (Van Hoof 2005). Ainsi, dans cet exemple, la préservation des deux paralogues ne peut pas être expliquée par un scénario de néo-fonctionnalisation. En revanche, cet exemple illustre le modèle de sous-fonctionnalisation, où deux paralogues partagent les différentes sous-fonctions (partenaires d'interaction, domaines d'expression) d'un gène pléiotrope ancestral, et sont ainsi préservés sur le long terme (figure 12c). Il est important de noter ici que deux chemins évolutifs peuvent aboutir à la sous-fonctionnalisation des deux copies. Le premier chemin, décrit par le modèle de *duplication, dégénération et complémentation*

(DDC), est un processus qui fait intervenir uniquement des mutations neutres (Force *et al.* 1999). Dans le modèle DDC, les deux copies évoluent sous les effets combinés de la sélection naturelle et de la dérive génétique (relâchement de la sélection négative). L'accumulation de mutations délétères complémentaires entre les copies aboutit à la perte réciproque (et donc au partage) des différentes sous-fonctions du gène ancestral. Les mutations restent en apparence neutres (bien que défavorables) dans la mesure où les fonctions perdues par une copie sont toujours réalisées par la seconde. La sous-fonctionnalisation par DDC est un modèle puissant pour expliquer la rétention en deux copies de nombreux gènes présents dans les génomes eucaryotes. Le second chemin évolutif, qui aboutit également à la sous-fonctionnalisation des deux copies, fait intervenir des mutations adaptatives, et est décrit par le modèle de *l'évitement du conflit adaptatif* (ECA) (Des Marais & Rausher 2008). Dans le modèle ECA, le gène ancestral possède deux sous-fonctions que la sélection naturelle ne peut simultanément optimiser (conflit adaptatif). Ainsi, après duplication, les deux copies peuvent contourner le conflit, chacune pouvant accumuler des mutations adaptatives, et se spécialiser respectivement dans une des deux sous-fonctions du gène ancestral. Finalement, ces deux modèles évolutifs (DDC et ECA), difficiles à distinguer, conduisent à la sous-fonctionnalisation des deux copies, et à leur préservation sur le long terme. Les deux paralogues *SIR3* et *ORC1* illustrent un cas clair de sous-fonctionnalisation, mais les données expérimentales ne permettent pas de distinguer entre les modèles DDC et ECA (Van Hoof 2005, Conant & Wolfe 2008).

La co-optation ou l'émergence de nouvelles fonctions

Les études engagées au cours des dernières décennies pour comprendre l'évolution des gènes dupliqués montrent que la plupart sont préservés dans les génomes eucaryotes par le mécanisme de sous-fonctionnalisation, ce qui ne suppose aucune émergence de nouvelles fonctions (Lynch 2007). Cependant, de nombreuses innovations évolutives identifiées au cours de l'histoire évolutive des eucaryotes sont la conséquence de duplications de gènes (et de génome) (Holland 2013). Pour comprendre ce paradoxe, il faut considérer que la lecture taxinomique des différents modèles ci-dessus, présentés séparément et sous la forme de trajectoires évolutives exclusives, est valide uniquement pour caractériser les mécanismes évolutifs impliqués dans la préservation des gènes dupliqués. Les gènes préservés en deux copies, par sous-fonctionnalisation, ou par effet de dosage par exemple, continuent de survivre sur de longues périodes évolutives,

augmentant les chances ultérieures de néo-fonctionnalisation (Bridgham *et al.* 2008). Une nouvelle fonction peut être alors la conséquence de changements dans la séquence codante, qui attribuent de nouvelles propriétés (interactions avec de nouveaux partenaires, nouveaux sites de régulation post-traductionnels, perte d'un domaine), et/ou de changements au niveau des séquences non-codantes (acquisition ou perte d'éléments *cis*-régulateurs notamment). Conant & Wolfe (2008) rappellent ainsi que la préservation sur le long terme des gènes dupliqués dans les génomes favorise l'apparition ultérieure de nouvelles fonctions, et ajoutent que ces nouvelles fonctions semblent apparaître principalement par co-optation de fonctions pré-existantes. Ils notent que l'éventail des fonctions d'un gène qui peut être co-optées, depuis son rôle majeur jusqu'aux activités mineures, est très large (figure 13). En effet, l'idée qu'une protéine catalyse une unique réaction dans la cellule, ou qu'un récepteur fixe un unique ligand, est une simplification très utile pour les études de biochimie, mais reconnue depuis longtemps comme très éloignée de la réalité (Jensen 1976). Les protéines possèdent de nombreuses activités mineures, parfois éloignées de leur fonction principale. Cette tolérance pour différents substrats (au sens large du terme) constitue une opportunité pour l'évolution de nouvelles fonctions par co-optation, en particulier après duplication de gène. Ainsi, le rôle majeur d'une protéine peut être recruté pour réaliser une nouvelle fonction, comme par exemple la co-optation, après duplication chez la levure, de la nouvelle copie de l'alcool déshydrogénase pour la consommation d'éthanol (Thomson *et al.* 2005). A l'inverse, certaines protéines semblent avoir été co-optées pour des fonctions très éloignées de leurs rôles initiaux, comme par exemple certaines enzymes du métabolisme, recrutées au niveau de la lentille de l'œil comme protéines de structure (cristallines) (Wistow & Piatigorsky 1987). Le chapitre suivant débute par une illustration de l'importance des duplications de génome pour l'innovation génétique, en prenant l'exemple des vertébrés. Cette partie montre combien la co-optation après duplication est un mécanisme majeur pour l'acquisition de nouvelles fonctions. En effet, l'élaboration des nouvelles structures embryonnaires uniques aux vertébrés à partir de structures préexistantes semble être associée à de nombreux événements de co-optation de nouveaux gènes issus des deux duplications de génome survenues à la base des vertébrés (Holland 2013).

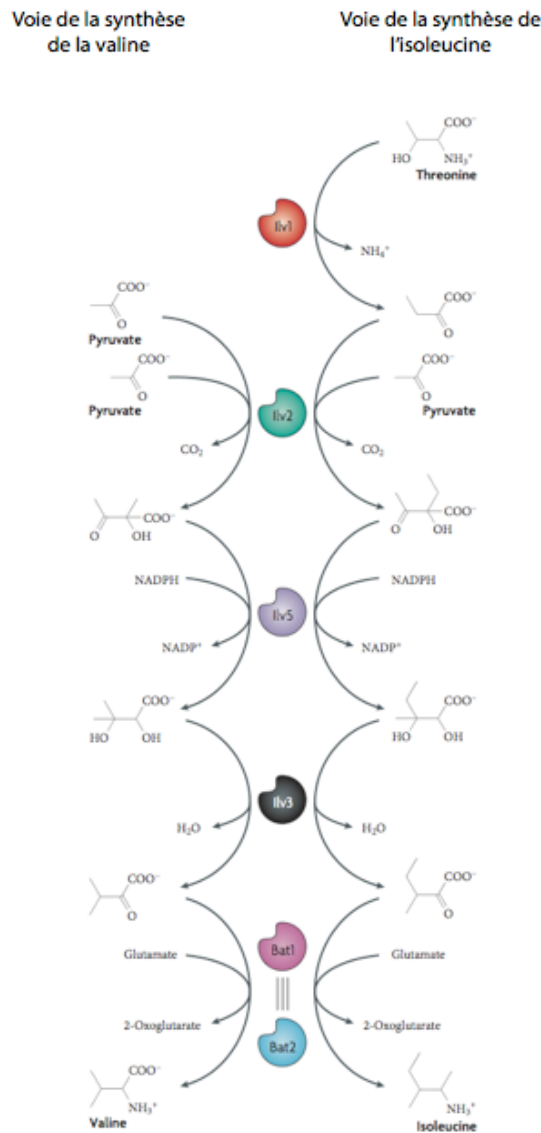


Figure 13: Voies de biosynthèse de la valine et de l'isoleucine chez la levure *Saccharomyces cerevisiae* (Conant & Wolfe 2008). Plusieurs enzymes au sein des ces deux voies métaboliques catalysent des réactions chimiques analogues à partir de substrats similaires mais distincts. Cet exemple illustre le fait que des enzymes possèdent plusieurs activités mineures, qui peuvent être co-optées pour de nouvelles fonctions au cours de l'évolution, notamment après duplication.

La portée évolutive des duplications de génome

Duplications de génome, origine et diversification des vertébrés

Les mécanismes génétiques qui sous-tendent l'origine évolutive des vertébrés, et les innovations évolutives majeures associées, intéressent les biologistes depuis de très nombreuses années. En 1970, le généticien Ohno propose que les événements de polyploïdisation survenus à la base des vertébrés soient impliqués dans l'origine évolutive et la diversification des vertébrés, notamment de par l'enrichissement du contenu génique des génomes, et ainsi les nombreuses opportunités de néofonctionnalisation. Un grand nombre de travaux engagés sous l'impulsion de cette hypothèse pionnière ont confirmé l'expansion et la diversification de nombreuses familles de gènes du développement (facteurs de transcription, protéines de signalisation), et des réseaux de régulation génétique associés, en particulier ceux qui gouvernent le développement des trois structures embryonnaires qui classiquement définissent les vertébrés : la crête neurale, les placodes neurogéniques, et un cerveau tripartite (Holland 2009, Holland 2013). Ces nouvelles structures participent ensemble, entre autres, au processus de céphalisation (avec le développement d'une tête) et sont certainement responsables de la transition vers un comportement de prédation active des animaux vertébrés (Gans & Northcutt 1983). Ainsi, actuellement, la majorité des scientifiques (mais pas la totalité, voir Donoghue & Purnell 2005) s'accorde pour attribuer aux duplications de génome un rôle majeur dans l'évolution des vertébrés, de par l'augmentation puis la diversification sur le long terme des répertoires de gènes du développement (Van de Peer *et al.* 2009, Cañestro *et al.* 2013). Cependant, des études récentes précisent ce rôle, et indiquent que les deux duplications de génome sont responsables plus de la diversification que de l'origine des innovations évolutives chez les vertébrés. En effet, ces travaux montrent que les trois structures embryonnaires mentionnées plus haut sont présentes à l'état rudimentaire chez les urochordés et donc certainement chez le dernier ancêtre commun des vertébrés et urochordés (Mazet *et al.* 2005, Holland 2009, Abitua *et al.* 2012, Cañestro *et al.* 2013). Ainsi, les deux duplications de génome semblent avoir facilité l'élaboration et la diversification de structures embryonnaires pré-existantes (et non pas non-existantes, selon Ohno 1970), notamment par co-optation de nouveaux gènes du développement au niveau des réseaux de

régulation génétique sous-jacents. L'évolution de la crête neurale chez les vertébrés illustre cette hypothèse.

De l'impact des duplications de génome sur l'évolution de la crête neurale chez les vertébrés

Chez les vertébrés, les cellules de la crête neurale sont une population transitoire de cellules ectodermiques embryonnaires multipotentes localisée dans la région dorsale du tube neural, qui, après délamination et migration vers différentes régions de l'embryon, se différencie en un très large éventail de dérivés mésenchymateux (os, cartilage notamment) et non-mésenchymateux (neurones, glie, et mélanocytes). Dans la mesure où de nombreux caractères qui différencient les vertébrés des autres chordés dérivent de la crête neurale (dont l'acquisition d'une tête, mentionné plus haut), cette structure embryonnaire a longtemps été considérée comme une innovation vertébrés-spécifique (Bronner & LeDouarin 2012). Cependant, les travaux de Levine et son équipe (2012) ont permis récemment d'identifier clairement chez un urochordé une structure embryonnaire homologue à la crête neurale des vertébrés. De plus, cette population cellulaire exprime un ensemble de gènes (*Id*, *Snail*, *Ets* et *FoxD*) homologues au réseau de régulation génétique impliqué dans la spécification des cellules de la crête neurale chez les vertébrés (Abitua *et al.* 2012). Les auteurs montrent également que ces cellules sont capables de se différencier en mélanocytes (dérivés non-mésenchymateux) mais qu'elles ne possèdent pas, en revanche, le réseau de régulation génétique (dont le gène *Twist*) qui gouverne la différenciation des cellules de la crête neurale en cellules mésenchymateuses chez les vertébrés. Enfin, ils montrent que l'expression ectopique du gène *Twist* dans cette population cellulaire est suffisante pour induire la migration et la différenciation en cellules mésenchymateuses (semblables aux cellules ecto-mésenchymateuses des vertébrés). L'ensemble de ces travaux suggère que le développement de toutes les potentialités des cellules de la crête neurale chez les vertébrés (en particulier leur différenciation en cellules ecto-mésenchymateuses) est la conséquence d'événements de co-optation de gènes (par exemple *Twist*) au niveau des réseaux de régulation génétique pré-existants, certainement facilités par l'enrichissement des répertoires de gènes au cours des deux événements de duplication de génome.

Duplications de génome et spéciation

L'étude des mécanismes d'isolement reproductif est abordée par les biologistes depuis de très nombreuses années, car la notion d'espèce et donc de spéciation est fondamentale pour comprendre la diversité des espèces à reproduction sexuée et leur évolution. En particulier, dans les années 1930-1940, Dobzhansky et Muller étudient les bases génétiques de l'isolement reproductif et proposent indépendamment un modèle pour expliquer l'origine des incompatibilités génomiques inter-spécifiques (Dobzhansky 1937, Muller 1940). Ce modèle, initialement décrit par Bateson (1909), et nommé le modèle de Bateson-Dobzhansky-Muller (BDM), postule que la divergence allélique de deux (ou plusieurs) loci avec épistasie (par dérive génétique ou évolution adaptative) entre deux populations allopatriques peut conduire à un isolement génétique de ces deux populations (avec dépression des hybrides). Ainsi, l'association de deux combinaisons alléliques, pour autant respectivement favorables dans les fonds génétiques des populations parentes, conduit à des effets épistatiques négatifs dans la population hybride (figure 14). Quelques décennies plus tard, Werth et Windham (1991) puis Lynch et Force (2000) proposent un modèle (compatible avec le modèle BDM) qui postule que la perte de différentes copies de gènes dans des populations séparées (par dérive génétique ou évolution adaptative) peut également aboutir à un isolement reproductif de ces populations (avec des hybrides sans aucune copie en F₂). Ainsi, les auteurs enrichissent le modèle BDM en considérant le polymorphisme structural en plus du polymorphisme de substitution nucléotidique dans le développement des incompatibilités génomiques interspécifiques. En particulier, la perte différentielle entre différentes populations de centaines de gènes issus d'une duplication de génome (par non-fonctionnalisation, voir chapitre précédent) peut faciliter la spéciation (Van de Peer *et al.* 2009). Par exemple, chez trois espèces de levures issues d'un ancêtre commun polyploïde, Wolfe et son équipe ont montré que, pour de très nombreux loci, des pertes différentielles sont survenues pour une même paire de gène, et ainsi 4% à 7% des singletons comparés entre espèces sont paralogues, et non pas orthologues (Scannell *et al.* 2006). Ces travaux illustrent le modèle de spéciation par perte réciproque de gène. Finalement, Lynch et Force (2000) ajoutent que la sous-fonctionnalisation des deux copies d'un gène après duplication constitue également un mécanisme génétique majeur de spéciation. En effet, les deux copies peuvent sous-fonctionnaliser différemment dans des populations géographiquement isolées, et les orthologues dans les différentes populations réalisent alors différentes fonctions. Après hybridation, en théorie, un

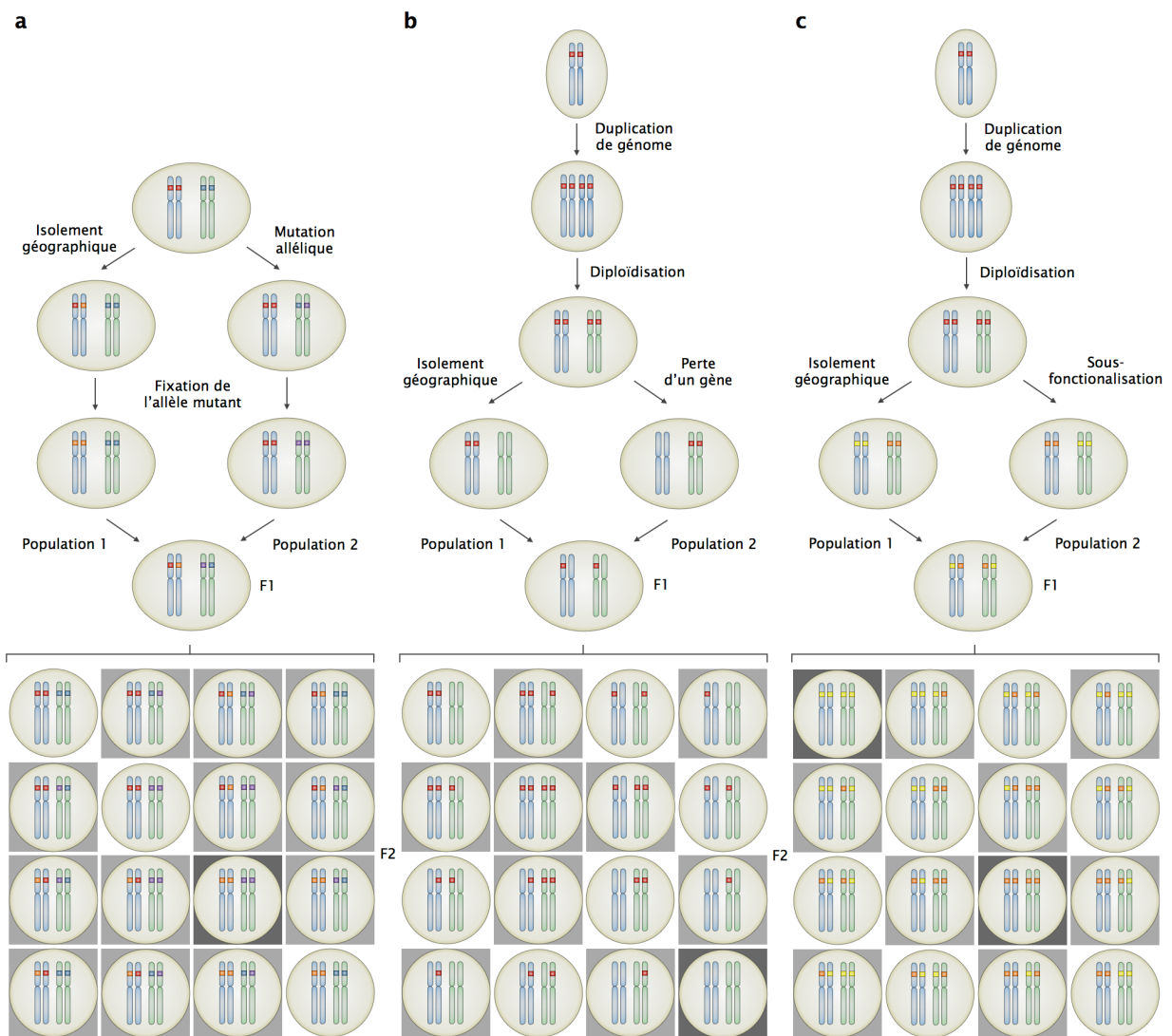


Figure 14: Mécanismes génétiques d'isolement reproductif. (a) *Modèle d'incompatibilité génétique de Bateson-Dobzhansky-Muller.* Les bandes rouges et bleues représentent deux loci avec épistasie sur deux chromosomes différents. Après isolement géographique, un allèle mutant se fixe au locus rouge dans la population 1 (bande orange), tandis qu'un allèle mutant se fixe au locus bleu dans la population 2 (bande violette). Le croisement des deux populations donne des individus hybrides (F₁) hétérozygotes pour chaque locus. Les allèles mutants sont favorables dans les génotypes parents mais leur combinaison dans le génotype hybride peut être délétère. Après croisement des F₁, un-seizième des individus F₂ sera homozygote pour l'allèle mutant aux deux loci épistatiques (carré gris foncé) et génétiquement incompatibles. Des incompatibilités génétiques sont également rencontrées pour 3/4 des autres individus F₂ (carrés gris). (b) *Non-fonctionnalisation différentielle après duplication.* Les bandes rouges sur les chromosomes représentent un locus (ici un gène) dupliqué au cours d'un événement de polyploïdisation. Après diploïdisation, les deux copies du gène sont présentes sur deux chromosomes différents. Après isolement géographique, les deux populations ont perdu une copie du gène sur un chromosome différent. Si les individus des deux populations se croisent, la population hybride sera hétérozygote, avec un unique allèle fonctionnel à chacun des deux loci. Un-seizième des individus F₂ recevra aucun allèle fonctionnel (carré gris foncé), et si ce gène est essentiel, sera non-viable. D'autres présenteront des problèmes d'haploinsuffisance ou de conflit de dosage (carrés gris). (c) *Sous-fonctionnalisation différentielle après duplication.* Dans ce cas, après diploïdisation et isolement géographique, les gènes dupliqués ont sous-fonctionnalisé différemment dans les deux populations (bandes orange ou jaune sur les chromosomes). La population hybride F₁ pourra se développer normalement, mais deux-seizième des F₂ recevra un seul allèle qui a perdu une sous-fonction essentielle (carrés gris foncé), diminuant la valeur sélective des individus. Comme précédemment, d'autres F₂ auront une valeur sélective réduite par des problèmes d'haploinsuffisance ou de conflit de dosage (carrés gris).

seizième des zygotes hybrides F2 sera ainsi dépourvu d'une sous-fonction, et mourra si cette fonction est essentielle (figure 14). Finalement, après duplication de génome, la non-fonctionnalisation et/ou la sous-fonctionnalisation différentielles de centaines de paires de gènes entre différentes populations peuvent promouvoir la spéciation sur de très longues périodes évolutives (Lynch & Force 2000, Sémon & Wolfe 2007a). Cependant, le développement de tels modèles doit prendre en considération le fait que les radiations évolutives chez les animaux ne sont pas toujours précédées d'un événement de polyploïdisation (Van de Peer *et al.* 2009).

De l'impact des duplications de génome sur la radiation des téléostéens

De nombreux auteurs proposent que la duplication de génome survenue chez l'ancêtre commun des téléostéens soit à l'origine des très nombreuses espèces qui composent ce clade (revu dans Volff 2005, Van de Peer *et al.* 2009). En effet, comme mentionné dans le premier chapitre, les poissons téléostéens constituent 47% des espèces de vertébrés actuels connus et 96% des espèces des actinoptérygiens (Lecointre *et al.* 2010). Néanmoins, une longue période de temps sépare l'événement de polyploïdisation (estimé entre 226 et 350 millions d'années) de la radiation évolutive des téléostéens survenue quelques 150 millions d'années plus tard, à la limite du Crétacé et du Tertiaire. Cependant, comme mentionné précédemment, les mécanismes de non-fonctionnalisation et sous-fonctionnalisation différentielles peuvent survenir des dizaines de millions d'années après duplication de génome, et promouvoir la spéciation sur de très longues périodes de temps, potentiellement facilitée par l'apparition de nouvelles niches écologiques suite à un bouleversement climatique tel que l'extinction Crétacé-Tertiaire par exemple (Van de Peer *et al.* 2009, Scannell *et al.* 2006, Sémon & Wolfe 2007a). Une étude réalisée chez deux poissons téléostéens, le poisson-zèbre et le tétraodon, montre que des pertes différentielles entre ces deux espèces sont survenues pour 8% des paires de gènes (1700 loci) issues de la duplication de génome survenue à la base des téléostéens (Sémon & Wolfe 2007b). De même, la comparaison des génomes des médaka et poisson-zèbre a permis d'identifier des cas de pertes différentielles (Naruse *et al.* 2004). Par ailleurs, des cas de sous-fonctionnalisation espèce-spécifique sont identifiés pour de nombreux gènes chez les poissons téléostéens, et ce quelques fois entre deux espèces très apparentées (Volff 2005). Ces travaux sont en accord avec l'idée que les événements de non-fonctionnalisation et sous-fonctionnalisation peuvent survenir après duplication de génome et promouvoir la spéciation sur de très longues périodes évolutives (Sémon & Wolfe 2007).

Duplication de génome, exaptation et complexification

Une duplication de génome est une opportunité pour moduler drastiquement le contenu génique d'un génome. En effet, les études de génomique comparative montrent que les duplications de génome augmentent sur le long terme le nombre de gènes dans les génomes eucaryotes (Lynch 2007). De plus, les mêmes études montrent que les duplications de génome enrichissent les génomes principalement en gènes du développement (qui sont sensibles à l'effet de dosage, voir chapitre précédent) dans des taxa aussi divers que les plantes, les vertébrés et les champignons (Van de Peer *et al.* 2009). Les biologistes Freeling et Thomas (2006) proposent que la sélection négative, pour préserver un équilibre du dosage, maintient mécaniquement en deux copies les gènes sensibles à l'effet de dosage, et ce de manière non-adaptative. Les gènes du développement et les unités fonctionnelles associées, ainsi préservés en deux copies, peuvent être secondairement recrutés sous l'effet de la sélection positive pour de nouvelles fonctions (qui seront en cela des exaptations). Sous cette hypothèse, les duplications de génome, en favorisant mécaniquement la rétention des gènes sensibles au dosage, peuvent conduire à certaines innovations évolutives sur le long terme, par exemple la céphalisation des vertébrés mentionnée plus haut, qui auraient eu peu de chance d'apparaître autrement (Conant & Wolfe 2008). Enfin, Freeling et Thomas proposent que les duplications de génome qui jalonnent l'évolution des eucaryotes, en repoussant les limites des phénotypes possibles (notamment de par la diversification spatio-temporelle des territoires présomptifs chez l'embryon), et en augmentant la complexité des génomes, sont responsables de la complexification morphologique des organismes multicellulaires. Cette hypothèse est malheureusement difficile, si ce n'est impossible, à tester (Sémon & Wolfe 2007a). Donoghue et Purnell (2005) critiquent cette dernière hypothèse et affirment que la prise en compte des lignées fossiles chez les vertébrés révèle qu'il n'y a pas eu de saut évolutif particulier (notamment en terme d'innovation morphologique) à la base des vertébrés, suite aux deux duplications de génome. Cependant, les unités fonctionnelles dupliquées constituent en quelque sorte, selon la métaphore de Gould (1979), des «pendentifs» dans le génome, qui peuvent être co-optés pour de nouvelles fonctions plus ou moins longtemps après l'événement de polypléidisation (Gould & Lewontin 1979). Par exemple, les opportunités de co-optation peuvent être facilitées par une augmentation du taux de spéciation, qui peut survenir longtemps après l'événement de polypléidisation, suite à un bouleversement environnemental par exemple (Van de Peer *et al.* 2009).

La famille des facteurs de transcription Sox

Dans les cellules eucaryotes, la transcription des gènes par les ARN polymérases nécessite le recrutement de différentes classes de protéines, dont les facteurs de transcription généraux, les co-facteurs, les histones, et les protéines chromosomiques non histones. En particulier, les facteurs de transcription, en ajustant l'expression de différents gènes cibles, assurent le bon fonctionnement des activités cellulaires, et permettent à la cellule de s'adapter aux changements de son milieu. Les facteurs de transcription, qui représentent moins de 10% des gènes codant des protéines, jouent cependant un rôle fondamental dans divers processus biologiques dont notamment le cycle cellulaire, le métabolisme, l'immunité, la reproduction, ou encore la prolifération et la différenciation des cellules au cours du développement (Vaquerizas *et al.* 2009). L'importance de cette classe de protéines dans la biologie de la cellule, et plus généralement des organismes vivants, explique la forte représentation des facteurs de transcription dans les maladies humaines (cancers et troubles du développement). Un important travail pour identifier et répertorier l'ensemble des facteurs de transcription présents dans le génome humain montre que 80% des facteurs de transcription contiennent soit un domaine doigt de zinc (motif C₂H₂), soit un homéodomaine, soit un domaine hélice-boucle-hélice (Vaquerizas *et al.* 2009). Cette même classification indique que la super-famille des protéines HMG (qui inclut la famille Sox, présentée ci-après) constitue la 8ème classe (sur environ 25 classes) la plus importante en nombre de facteurs de transcription dans le génome humain. Etant donnée l'importance des différentes familles de facteurs de transcription dans le développement et la physiologie des animaux, étudier les mécanismes évolutifs impliqués dans l'enrichissement et la diversification fonctionnelle de ces familles de gènes est fondamental pour comprendre les bases génétiques de la diversification phénotypique des animaux. Le présent manuscrit se concentre sur la famille des facteurs de transcription Sox retrouvée chez tous les métazoaires.

Découverte du premier membre de la famille Sox

La découverte simultanée du gène du déterminisme du sexe chez l'homme et la souris en 1990 a conduit à l'identification du premier membre de la famille des gènes Sox, localisé sur le bras court du chromosome Y, et nommé *Sry* (*Sex-determining region of Y*

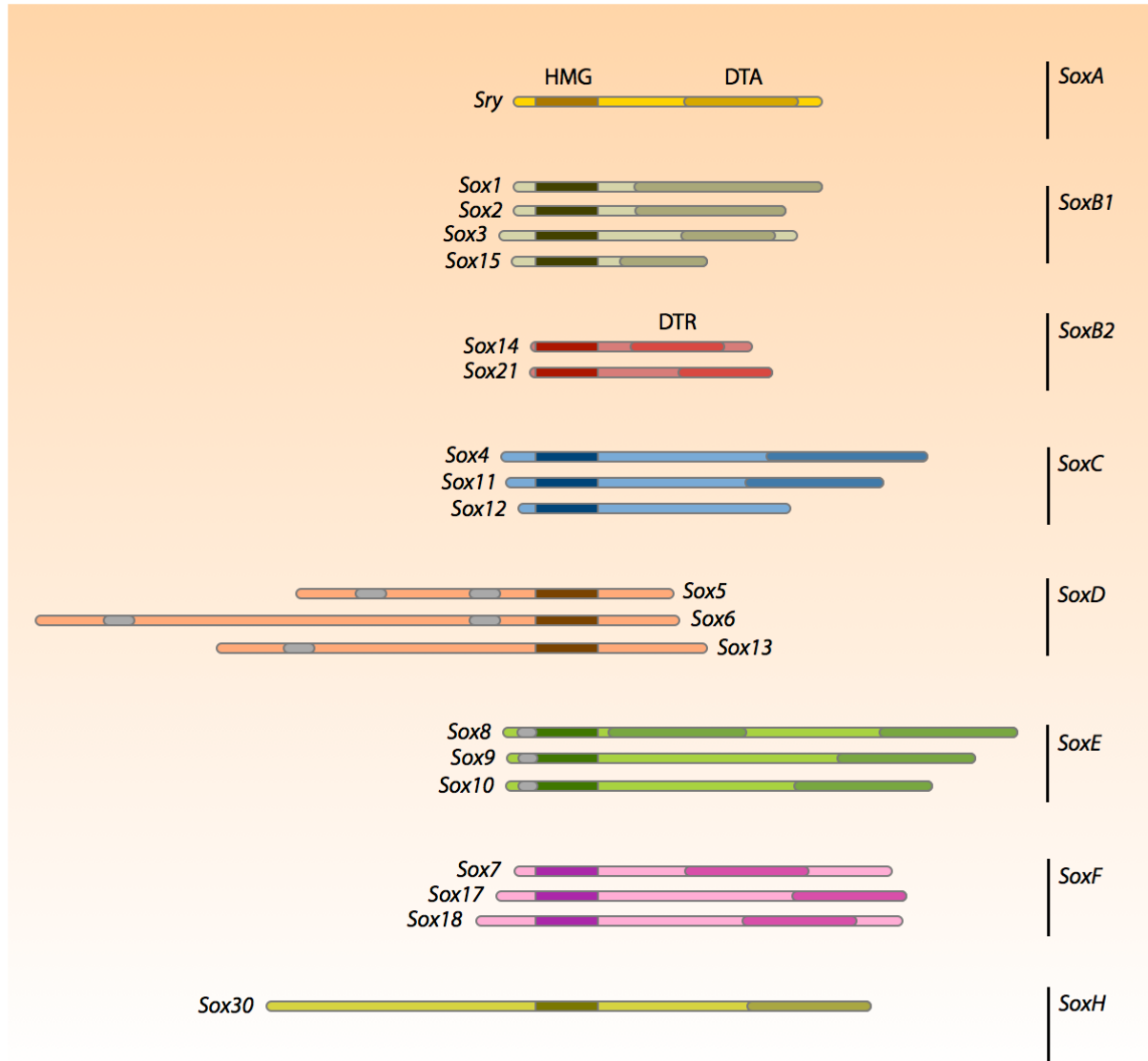


Figure 15: Classification des membres de la famille Sox (d'après Lefebvre et al. 2007). La boîte HMG est représentée par un rectangle, tandis que les domaines de transactivation DTA (trans-répression DTR pour les SoxB2) sont représentés par des cylindres. Les domaines de dimérisation sont représentés par des cylindres gris.

chromosome) (Sinclair *et al.* 1990, Gubbay *et al.* 1990). Comme mentionné dans le paragraphe précédent, les facteurs de transcription Sox appartiennent à la super-famille des protéines HMG (Ner 1992, Wegner 1999) qui toutes possèdent au moins un domaine HMG (*high-mobility group*, pour rappeler le comportement de ces protéines durant une séparation électrophorétique) de fixation à l'ADN. Cette grande famille se subdivise en deux groupes dont le premier contient des protéines chromosomiques non histones HMG qui interviennent dans l'architecture chromatinienne. Le second groupe est composé de deux familles de facteurs de transcription : la famille Lef/Tcf (*lymphoid enhancer factor/T-cell factor*), et la famille Sox (*SRY-related HMG box* ; Bowles *et al.* 2000). Cependant, les facteurs Lef/Tcf sont des facteurs de transcription qui ne possèdent pas de domaine de transactivation. C'est le complexe bêta-caténine/Lef-Tcf qui reconstitue un facteur de transcription fonctionnel, notamment au niveau de la voie Wnt (*wingless-related integration site*)/bêta-caténine. En revanche, les facteurs de transcription Sox sont capables d'interagir avec d'autres protéines par l'intermédiaire d'un domaine de transactivation et de réguler ainsi la transcription de gènes cibles (Sarkar & Hochedlinger 2013). Les résidus spécifiques du domaine HMG de la protéine SRY ont permis au début des années 1990 la rapide identification des autres membres de la famille Sox localisés sur d'autres chromosomes de la souris et de l'homme, et également dans les génomes de différentes autres espèces vertébrés et invertébrés (Gubbay *et al.* 1990, Denny *et al.* 1992, Coriat *et al.* 1993). Chez les mammifères, vingt gènes apparentés à *Sry* ont été identifiés, et constituent les vingt membres de la famille des gènes Sox repartis en huit groupes (A à H) (figure 15).

Evolution et diversité fonctionnelle de la famille Sox

Les gènes Sox constituent une ancienne famille de facteur de transcription qui trouve son origine chez les unicellulaires eucaryotes (King *et al.* 2008). L'émergence des métazoaires est associée à une expansion de la famille Sox survenue très probablement par événements de duplication et/ou brassage de domaines d'un unique gène ancestral puis divergence des différentes nouvelles copies : l'ancêtre commun des bilatériens avait très certainement sept gènes Sox (*B1, B2, C, D, E, F* et *H* ; Larroux *et al.* 2006, Larroux *et al.* 2008, Bowles *et al.* 2000). En accord avec cette hypothèse, tous les génomes bilatériens contiennent au moins ces sept gènes Sox (Bowles *et al.* 2000, Phochanukul & Russell 2010, Paps *et al.* 2012). Les animaux vertébrés possèdent plusieurs copies de chaque gène très probablement produites au cours des deux duplications de génome

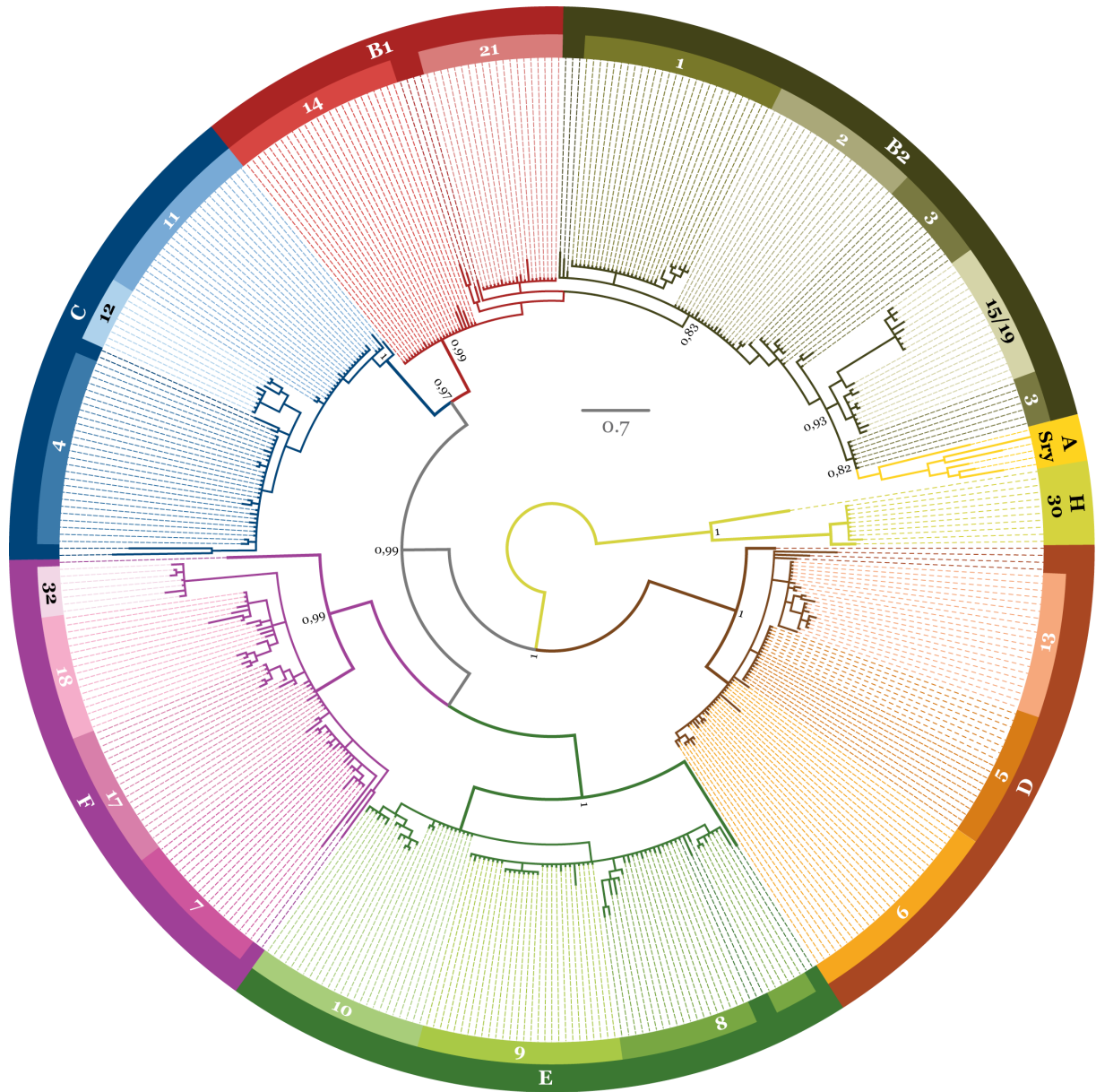


Figure 16: Phylogénie de la famille Sox. La reconstruction phylogénétique par inférence bayésienne est basée sur la séquence protéique de la boîte HMG des protéines Sox. Chaque groupe contient au moins un représentant des invertébrés (figure réalisée pour la revue en préparation, en annexe de ce manuscrit).

survenues à la base des vertébrés (Okuda *et al.* 2006). Ainsi, la famille de gènes *Sox* compte chez les vertébrés 19 membres répartis en 7 groupes (B1 à H, figure 16). Les mammifères possèdent un membre supplémentaire, le gène *Sry* (qui constitue le groupe A) présent sur le chromosome Y, qui résulte de la divergence des deux allèles *Sox3* survenue au cours de la dégradation du chromosome Y chez les mammifères (l'unique allèle *Sox3* est localisé chez les mammifères sur le chromosome X) (Graves 2006). Les gènes *Sox* jouent un rôle fondamental dans le développement et l'homéostasie des vertébrés, et plus généralement des animaux bilatériens. Ils participent à une très importante diversité de processus biologiques, depuis la reprogrammation du génome du zygote (acquisition de la totipotence), à la différenciation des cellules souches embryonnaires et adultes au sein de très nombreux tissus, depuis la spécification des cellules de la crête neurale, à la formation des os du crâne (Guth & Wagner 2008, Sarkar & Hochedlinger 2013). Le présent manuscrit présente brièvement dans cette introduction le rôle des *Sox* dans la neurogénèse et la biologie de la reproduction chez les vertébrés pour illustrer les modes d'actions et les fonctions des facteurs *Sox* dans la biologie des animaux. Une revue en préparation, en annexe de ce manuscrit, reprend et développe les différents aspects abordés dans cette introduction sur l'évolution et le rôle biologique de la famille *Sox* chez les vertébrés.

Les facteurs Sox dans la neurogénèse : facteurs de transcription et facteurs pionniers

Le système nerveux central des vertébrés se développe à partir des cellules souches neuronales, qui deviennent des cellules précurseurs neuronales, prolifèrent dans le neuro-épithélium, et finalement se différencient en cellules neuronales ou gliales. Les différentes étapes du neurodéveloppement impliquent successivement plusieurs facteurs de transcription *Sox*, en particulier les *Sox* des groupes B1 et C (figure 15). Ainsi, le facteur *Sox2*, exprimé dans les cellules souches, est rejoint par les deux *Sox1* et *Sox3* dans les cellules précurseurs, et ces trois facteurs *SoxB1* sont relayés par *Sox4* et *Sox11* (*SoxC*) lorsque les cellules s'engagent dans la voie de différenciation (Guth & Wegner 2008, Bergsland *et al.* 2011, Wegner 2011). Au cours du neurodéveloppement, les facteurs *Sox* agissent comme des facteurs de transcription, et activent de nombreux gènes cibles, mais également comme des facteurs pionniers, qui prépare l'environnement chromatinien au niveau des régions amplificatrices (ou «enhancer») pour la fixation ultérieure d'autres facteurs de transcription (Wegner 2011). Par exemple, dans les cellules précurseurs neuronales, les facteurs *SoxB1* se positionnent en amont des gènes cibles pour initier leur expression, mais également en amont de gènes

qui seront activés uniquement lorsque les cellules s'engageront dans la voie de différenciation. En effet, les SoxB1, qui auront préparé un environnement chromatinien favorable à l'activation des gènes impliqués dans la différenciation, seront relayés au niveau de ces amplificateurs par les SoxC qui joueront quant à eux le rôle de facteurs de transcription (Bergsland *et al.* 2011). Il est intéressant de constater que les fonctions des SoxB1 et SoxC au cours du neurodéveloppement sont maintenues dans la neurogénèse adulte chez les mammifères (Lefebvre *et al.* 2007, Mu *et al.* 2012). Finalement, les facteurs pionniers sont depuis peu considérés comme des facteurs clés dans le développement, l'homéostasie, et le développement des tumeurs chez l'homme, et par conséquent les facteurs Sox constituent un modèle d'étude pertinent pour fournir de nouvelles indications sur le mode d'action des facteurs pionniers dans ces processus biologiques.

Facteurs Sox et biologie de la reproduction

La famille Sox est également connue pour être impliquée dans la biologie de la reproduction des animaux, notamment après la découverte du célèbre gène *Sry* impliqué dans le déterminisme du sexe chez les mammifères (Guth & Wegner 2008). Le mode d'action du gène *Sry* pour induire le développement des testicules plutôt que des ovaires chez les mammifères est encore peu compris. Contrairement aux autres membres de la famille Sox qui sont largement exprimés dans le temps et dans l'espace au cours du développement, le gène *Sry* est transitoirement exprimé pendant quelques heures dans un unique tissu : les cellules somatiques des gonades. Pendant cette courte période de temps, la protéine *Sry* agit en synergie avec le récepteur nucléaire orphelin Sf1, cible un amplificateur spécifique du gène *Sox9*, et induit la différenciation des cellules somatiques gonadiques en cellules de Sertoli (Sekido & Lovell-Badge 2008). La famille des facteurs de transcription Sox joue également un rôle conservé chez les animaux dans la différenciation sexuelle. Ainsi, le singleton *SoxE* chez la drosophile (nommé *Sox100B*) est nécessaire pour la différenciation des testicules, tout comme son orthologue *Sox9* chez les vertébrés (Schartl 2004, Nanda *et al.* 2009). Chez les poissons téléostéens, de nombreux gènes *sox* sont exprimés dans les gonades mâle et femelle (Santos *et al.* 2007, Navarro-Martin *et al.* 2009, Navarro-Martin *et al.* 2012). Certains auteurs suggèrent que les différents gènes *Sox* exprimés dans les ovaires des poissons téléostéens peuvent avoir participé à l'élaboration du phénotype pélagique des œufs téléostéens (Zapater *et al.* 2013).

RÉSULTATS

Partie I. Expansion et évolution de la famille Sox chez les poissons téléostéens

L'impact des duplications de génome sur la biodiversité des espèces vertébrés reste une question ouverte et non résolue. En effet, les différentes hypothèses moléculaires avancées par les biologistes pour relier duplication de génome et diversification des espèces vertébrés sont souvent difficiles à tester. Par exemple, l'impact des deux duplications de génome survenues à la base des vertébrés sur l'enrichissement des répertoires de gènes du développement des vertébrés, bien que très probable, reste néanmoins difficile à démontrer pour de nombreuses familles de gènes du fait de l'ancienneté des duplications de génome (Dehal & Boore 2005). Chez les téléostéens, des études montrent qu'il est possible de déterminer clairement la contribution de la duplication de génome dans l'expansion d'une famille de gène, ou dans l'enrichissement d'une voie métabolique, dans la mesure où les poissons téléostéens partagent un ancêtre commun polyploïde plus récent que celui des vertébrés (Braasch *et al.* 2009, Ahn *et al.* 2012, Opazo *et al.* 2013). Ainsi, le groupe des téléostéens, qui compte plus de la moitié des espèces vertébrés existantes, constitue un modèle pertinent pour évaluer la contribution des duplications de génome dans l'expansion des familles multigéniques chez les vertébrés, pour comprendre les mécanismes évolutifs qui façonnent l'évolution des gènes et des familles de gènes, et finalement tester les hypothèses moléculaires qui relient duplication de génome et biodiversité.

La présente étude vise à déterminer l'impact de la duplication de génome survenue à la base des téléostéens sur l'évolution de la famille des gènes *sox* chez les poissons téléostéens. Dans un premier temps, cette étude a pour objectif (i) de récapituler le paysage des gènes *sox* dans 15 génomes de vertébrés (dont 10 génomes de poissons téléostéens) pour évaluer l'ampleur et le caractère spécifique de l'expansion de la famille des gènes *sox* chez les téléostéens, et (ii) de déterminer la contribution de la duplication de génome survenue à la base des téléostéens dans cette expansion. Les données de séquençage à disposition pour les 15 génomes étudiés, et les différentes banques de gènes et d'expression sont utilisées pour obtenir les paysages des gènes *sox* pour chaque espèce analysée. Les quatre génomes des poisson-zèbre, médaka, tétraodon et épinoche sont exploités pour réaliser une étude macrosyténique des différents gènes *sox* et évaluer l'impact de la duplication de génome dans l'expansion de cette famille

chez les téléostéens. Le second objectif de cette étude est de décrire l'évolution moléculaire des gènes *sox* dupliqués produits par la duplication de génome spécifique aux téléostéens, aussi bien au niveau de leurs séquences codantes que de leurs séquences non-codantes, pour inférer les mécanismes évolutifs principaux impliqués dans leur préservation, le cas échéant, sur le long terme. En particulier, les effets de la sélection naturelle sur l'évolution des séquences codantes des paralogues sont caractérisés en comparant les patrons de variations synonymes et non-synonymes (test de différents modèles d'évolution à codons), et la dynamique évolutive des séquences non-codantes environnantes est étudiée en recherchant les éléments très conservés entre les différentes espèces vertébrés. Enfin, les profils d'expression de 5 paires de gènes *sox* sont analysés chez trois espèces de poissons téléostéens pour obtenir des indications sur l'évolution de l'expression (et donc éventuellement de la fonction) des paralogues *sox* chez les téléostéens. Ces données sont destinées à améliorer notre compréhension du rôle des duplications de génome dans l'évolution d'une famille de gène fondamentale pour la biologie des vertébrés.

EXPANSION BY WHOLE GENOME DUPLICATION AND DIVERSIFICATION OF THE SOX FAMILY IN TELEOSTS

Emilien Voldoire, Frédéric Brunet, Magali Naville, Jean-Nicolas Volff, Delphine Galiana

Institut de Génomique Fonctionnelle de Lyon, UMR5242, Université de Lyon1, CNRS, INRA, Ecole Normale supérieure de Lyon. 32-34 avenue Tony Garnier 69007 Lyon.

Abstract

Several events of whole genome duplication (WGD) took place during vertebrate evolution. The fact that whole genome duplications are implicated in species creation, expansion and biodiversity is still a debated question, mainly due to the lack of experimental data. Indeed, the role of whole genome duplications in speciation and biodiversity is likely to be driven by the fate of the so-called duplicated genes created just after such an event: duplicates are in majority rapidly eliminated through non-functionalization mechanism. However, few of them (around 20%) are conserved in genomes and can be used as a new reservoir for genetic innovation through two processes: sub-functionalization and/or neo-functionalization. In this study, we have decided to analyze the impact of the teleost-specific whole genome duplication, which occurred around 250 to 350 mya, thanks to the analysis of the *sox* genes family evolution in teleostean fishes. Whole genome bioinformatics and large scale expression analyses have been conducted, leading to the conclusion that *sox* genes evolution in teleost have been impacted by the fish-specific whole genome duplication, and it seems that lineage-specific duplicates evolution is the main process by which this WGD event could have been implicated in teleost radiation.

Introduction

Origin and evolution of biodiversity remain fundamental questions in biology (Santos & Salzburger 2012). At the genetic level, the importance of gene duplication for innovation and diversification is recognized by geneticists since several decades (Haldane 1933, Ohno 1970, Kaessmann 2010). Especially, the duplication of all genes through whole-genome duplication (WGD), that occurred during a polyploidization event, provides a particularly remarkable opportunity for evolutionary novelties and functional diversification (Van de Peer *et al.* 2009, Freeling & Thomas 2006). Although evolutionary successful (*i.e.* retained) WGDs are rare (Van de Peer *et al.* 2009), several rounds of ancient WGD are inferred in vertebrates, including two rounds of WGD at the base of the vertebrate lineage (Ohno 1970, Dehal & Boore 2005). Moreover, it is now well-established that an additional WGD, named the teleost-specific WGD, have occurred in the lineage leading to the teleostean fishes about 250 mya (Amores *et al.* 1998, Jaillon *et al.* 2004, Kasahara *et al.* 2007). Furthermore, more recent and lineage-specific WGDs are inferred in various vertebrate lineages, such as in salmonids and in xenopus (Otto 2007).

Initially, a polyploidization event doubles the genome content and therefore the genetic content. Nonetheless, polyploidization is evolutionarily unstable and followed by diploidization (whereby tetraploid species become diploid) that reduce the genetic content (Wolfe 2001). Indeed, this process involves many genomic rearrangements and many gene deletion events. In addition, the two gene copies produced by WGD are redundant, and usually, one copy is free to accumulate degenerative mutations in its protein-coding sequence and/or its regulatory elements, as well as undergoing epigenetic modifications, and, by a process called pseudogenization, become lost from the genome (Lynch & Conery 2000). Thus, genomic studies in various taxa (yeast, plant, vertebrates) show that only 10% to 30% of duplicated genes are preserved after a WGD event (Kellis *et al.* 2004, Bowers *et al.* 2003, Brunet *et al.* 2006). However, these observations suggest the existence of mechanisms of preservation that counteract or exploit mutation-induced degeneration and protect from deletion events. It is proposed that, just after a WGD event, selection for dosage balance preservation represent an important force for the duplicate retention of dosage-balanced genes (the dosage balance hypothesis; Papp *et al.* 2003). This mechanism could explain the over-retention

of particular functional categories after WGD, often genes involved in complex molecular networks, such as developmental and transcription factor genes (Brunet *et al.* 2006, Aury *et al.* 2006, Kassahn *et al.* 2009). Other classic models for duplicated genes retention (and their expression retention) are sub-functionalization, in which the ancestral function is distributed between the two copies, and/or neo-functionalization, in which at least one of the two copies acquires a new function (Conant & Wolfe 2008). These models initially concern *cis*-regulatory elements, implying changes in gene expression (Sémon & Wolfe 2007), and protein product, involving changes in partner interactions (Lynch & Conery 2000). Interestingly, molecular signature into the protein-coding sequence, in terms of non-synonymous (dN) and synonymous substitution (dS) rates (*i.e.* the dN/dS ratio), provide an opportunity to infer the mechanism of preservation. Indeed, different predictions are made about the rate of sequence evolution of duplicates for neo-functionalization and sub-functionalization models (Chain & Evans 2006).

The fact that the two rounds of WGD occurred at the base of vertebrates coincide with the emergence of key evolutionary innovations in vertebrates, and the fact that the teleost-specific WGD gave rise to the most species-rich group in vertebrates, together suggest that WGDs provide opportunities for evolutionary novelties and can potentially facilitate diversification and speciation (Freeling & Thomas 2006, Van de Peer *et al.* 2009, Holland 2013). Indeed, at the genetic level, WGDs contribute to the expansion of regulatory gene classes (the over-retained categories mentioned above), and their subsequent functional diversification (in both their regulation and interaction partners) by subfunctionalization and/or neofunctionalization. Moreover, differential loss of gene duplicates and/or reciprocal gene loss in isolated populations (in the course of diploidization and by pseudogenization) can promote genomic incompatibilities and then speciation (Lynch & Conery 2000, Volff 2005, Sémon & Wolfe 2007).

In this study, we have analyzed the impact of the teleost-specific WGD on the evolution of the transcription factors Sox family in teleostean fishes. The SRY-related HMG box (or Sox) gene family, which originated in unicellular eucaryotes (King *et al.* 2008), has expanded over time in a lineage-specific manner through both small-scale duplication (SSD) and WGD events. In mammals, there are twenty members (Bowles *et al.* 2000) divided into eight groups (A, B1, B2, C, D, E, F, and H). This ancient family of transcription factors possesses fundamental functions in animal biology, especially

during embryogenesis, but also in adult homeostasis (Sarkar & Hochedlinger 2013). The famous mammal-specific *Sry* gene (group A), the first member identified approximately 20 years ago (Sinclair et al. 1990, Gubbay et al. 1990), is an excellent example of the Sox family importance : it initiates male sex determination during mammal development (Hawkins 1993). The seven other groups mentioned above are represented in all bilaterians, generally by a single *Sox* gene *per* group in invertebrates (Phochanukul et al. 2010). However, vertebrate genomes contain, for most groups, two or three *Sox* genes, as a probable consequence of the two rounds of WGD that occurred in the stem vertebrates (Okuda et al. 2006, Van de Peer et al. 2009). In addition, many studies show that several vertebrate *Sox* genes are duplicated in different teleostean genomes, and especially in the zebrafish model organism (De Martino et al. 2000, Chiang et al. 2001, Cresko et al. 2003, Hoegg et al. 2004, Kluever et al. 2005, Okuda et al. 2006, Mavropoulos et al. 2005, Zhang et al. 2008, Braasch et al. 2009, Zhang et al. 2010). These studies suggest that most *sox* duplicates in teleosts were preserved by subfunctionalization, often in a lineage-specific manner (De Martino et al. 2000, Cresko et al. 2003, Kluever et al. 2005).

In this context, the aim of this work was first to measure in what extent the Sox family has expanded in the teleost lineage, *id est* to draw by comparative genomics the landscape of all Sox members in all available teleost genomes, and therefore to determine the contribution of the teleost-specific WGD to the Sox family expansion within this lineage. Secondly, we investigated the mechanisms of preservation of some retained *sox* duplicates in several teleost genomes, by analyzing their genomic organization, and their expression pattern.

Materials and methods

Sequence retrieval and alignment

Nucleotide sequences of *Sox* genes were obtained using TBlastN searches against GenBank and Trace Archives Nucleotide from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), Harvard Expressed Sequence Tags bank (<http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/Blast/index.cgi>), and Ensembl genome assemblies (www.ensembl.org) of the mouse (*Mus musculus*), the coelacanth (*Latimeria chalumnae*), the spotted-gar (*Lepisosteus oculatus*), and the eight teleosts zebrafish

(*Danio rerio*), Cod (*Gadus morhua*), tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), platyfish (*Xiphophorus maculatus*), tetraodon (*Tetraodon nigroviridis*), fugu (*Takifugu rubripes*) and stickleback (*Gasterosteus aculeatus*). Sox sequences from the salmon (*Salmo salar*) and the elephant shark (*Callorhynchus milii*) were respectively obtained at <http://web.uvic.ca/grasp/gils/> and <http://esharkgenome.imcb.a-star.edu.sg/>. Sox genes and their accession numbers and/or coordinates are available in Supplementary data. Nucleotide sequences were loaded into MEGA (Tamura et al. 2007), translated into proteins and aligned using ClustalW (Tamura et al. 2007). The alignment were further refined manually using SeaView (Galtier et al. 1996; version 3.2).

Phylogenetic reconstructions

Maximum likelihood phylogenies were computed based on protein sequences with PhyML implemented in Seaview (Galtier et al. 1996; version 3.2). LG model of protein evolution was used and topology optimization was carried out using best of NNI and SPR option. Default aLRT (SH-like) were used for branch support (Anisimova & Gascuel 2006).

Synteny analyses

Macrosynteny analyses were performed on the four zebrafish *Danio rerio*, medaka *Oryzias latipes*, stickleback *Gasterosteus aculeatus* and tetraodon *Tetraodon nigroviridis* using their last release of genome assemblies in Ensembl. For each genome, a rose window representation was generated as in Jaillon et al. 2004, Brunet et al. 2006 and Braasch et al. 2009. Similarly, orthology relationships between chromosomes and between genes have been identified between the four studied species. To note, only 21 groups and the largest scaffold (scaffold_27) were used to generate the stickleback rose-window representation.

Search for conserved non-coding elements in sox regions

Two complementary methods were applied to look for conserved non-coding elements (CNEs) in the vicinity of fish Sox genes, either present up to human or being fish-specific. First, non-coding elements conserved at a large scale in vertebrates, and particularly found both in human and fish genomes, were identified in the 46-species multiple alignment obtained from the UCSC (University of California, Santa Cruz) using an algorithm called « ScanMaf ». This alignment include the following fish species: zebrafish (danRer7), medaka (otyLat2), stickleback (gasAcu1), Tetraodon (tetNig2) and fugu (fr2).

Basically, ScanMaf scans the alignment and looks for conserved regions of a minimal length (10bp) and identity (90 % in the 10bp-seed region, further extended by accepting up to 3 non-conserved columns each side). The program allows substitutions to occur, under a threshold of 12 %, in each column of the alignment. This relaxed definition of conservation allows thus to retrieve even loosely conserved regions, that could be missed by a simple homology search. It was applied apart from any exon and human masked sequences (rmk or trf), in 1Mb regions upstream and downstream SOX genes in human. Elements found up to at least one fish were retained. Elements shorter than 20bp, less than 35 % identical to the human sequence or closer than 20bp from any exon were finally discarded. As the multiple alignment does not take into account the fish-specific whole genome duplication, i.e. as it aligns any given human region with only one of its putatively two fish counterparts (the best matching one), previously identified CNEs were searched in paralogous SOX regions using BLAST. The same procedure was applied to look for them in three additional genomes : mouse (mm9), tilapia (oreNil1) and platyfish (xipMac1). Second, ray-finned fish-specific CNEs were searched in 800kb regions centered on the different *sox* genes using BLAST with the spotted-gar region, cleaned from any exonic and UTR elements, as query against 7 other fish Sox regions (zebrafish, medaka, tilapia, platyfish, Tetraodon and fugu), with following parameters : gap opening 2, gap extension 2, mismatch penalty -1, match reward 1. Only BLAST hits longer than 50bp and at least 60 % identical to the spotted gar sequence were retained. Regions of the spotted gar matching in different species and partially overlapping were fused in a larger resulting element. Regions with multiple matches, corresponding to repeated elements, were discarded.

Evolutionary rate computations

Alternative models with different branch-specific dN/dS ratio have been compared using the codeml program of PAML (Yang, 2007). Data set includes the two mammals mouse *Mus musculus* and human *Homo sapiens*, the chicken *Gallus gallus* and at least the five species of teleost fishes zebrafish *Danio rerio*, medaka *Oryzias latipes*, platyfish *Xiphophorus maculatus*, tilapia *Oreochromis niloticus* and fugu *Takifugu rubripes*. When available, sequences from the two non-teleost fishes spotted-gar *Lepisosteus oculatus* and coelacanth *Latimeria chalumnae*, the frog *Xenopus tropicalis*, and the six teleost fishes salmon *Salmo salar*, tetraodon *Tetraodon nigroviridis*, cod *Gadus morhua*, stickleback *Gasterosteus aculeatus*, orange-spotted grouper *Epinephelus coioides* and african sharptooth *Clarias gariepinus* were added. Alignment based on protein

sequences were executed with ClustalW implemented in MEGA (Tamura *et al.* 2007) and computation were done on the alignments after removing positions with at least one gap. As in Chain and Evans (2006), two analyses that test specific predictions about the dN/dS ratio were performed to assigned mechanisms for duplicate gene retention. First, the model A, in which a single dN/dS ratio is estimated for all branch, was compared with the model B, in which a higher dN/dS ratio is estimated after duplication, to test whether paralogues experienced relaxed selection after duplication. Secondly, the model B was compared with the model C, in which three dN/dS ratio are estimated, to test whether paralogues evolve differently from each other. Tests were done by comparing twice the difference of likelihood values to a X^2 distribution with degrees of freedom equal to the difference in number of free parameters of each model.

Fish

Platyfish (*Xiphophorus maculatus*, strain Usumacinta), zebrafish (*Danio rerio*, strain AB/TU) and medaka (*Oryzias latipes*, strain Hd-rR) were kept under standard conditions at the PRECI aquarium facility of the IFR128 Biosciences Gerland-Lyon Sud (Lyon, France). Fish embryos were raised to the required stages of development in E3 embryo media (Nüsslein-Volhard & Dahm 2002) at 26°C for medaka and 28,5°C for zebrafish. Medaka and zebrafish embryos were staged according to Iwamatsu (2004) and Kimmel *et al.* (1995), respectively. Mouse samples were provided by R. Guyot (Institute of functional genomic of Lyon, France). Experimental research reported here has been performed following internationally recognized guidelines (DDSV, Direction Départementale des Services Vétérinaires de Lyon, agreement number A 69 387 0602).

RT-qPCR expression analysis

Total RNA was extracted from separated male and female tissues (spleen, liver, ovaries, testes, muscle, eyes, brain) and from whole embryos using TRIReagent (Molecular Research Center, Inc.). After DNase treatment, four micrograms of total RNA were reverse-transcribed with the RevertAid First Strand Synthesis kit (Fermentas) using random hexamer primers. Real-time PCR was done on 2 µl of cDNA starting from a dilution of factor 15 using IQTM Custom SYBR® Green Supermix (Bio-rad). PCR amplification was monitored with a CFX96 RealtimeSystem (Bio-rad). Cycling conditions were 95°C for 2 min, 40 cycles of 95°C for 10 seconds and 59°C for 30 seconds. Three technical replicates were performed *per* sample. Quantification cycle (Cq; standard name for Ct or Cp value) values were recorded with the BioRad CFX manager version 2

software. Cq values ≥ 34 were considered beyond the limit of detection (a Cq value of 33 represents a single molecule template detection). *bActin2* and *rpl7* (ribosomal protein 17) were found to be the most stable reference genes respectively for embryos and adults, and used to normalize the data. Data were analyzed by the $\Delta\Delta CT$ method (Livak & Schmittgen 2001).

Results

The sox landscape in teleostean genomes

Genome-wide studies have been performed to provide the landscape of *sox* genes in the ten teleost genomes the most complete and adept for this large-scale analysis. The same work has been done in the five non-teleost species elephant shark (*Callorhinchus milii*), mouse (*Mus musculus*), human (*Homo sapiens*), coelacanth (*Latimeria chalumnae*) and spotted-gar (*Lepisosteus oculatus*) in order to have an idea of the evolutionary dynamics of this gene family in other vertebrate lineages. As a proxy to search for *sox* genes in annotated and non-annotated genomes, we used the sequences of the human *SOX*, except for the *SRY* gene (group A) which is mammal-specific. Phylogeny analysis have been done to determine if an identified sequence was truly related to a *sox* gene, and then to determine if it is a new teleost-specific *sox* gene or a duplicate of an already characterized *sox* gene in vertebrates (supplementary figure 1, A1 to K1). Results obtained are presented in figure 1. The closely related fish to teleosts, that does not experienced the teleost-specific WGD, the spotted-gar (Amores *et al.* 2011), contains no more than nineteen *sox* genes in its genome. These nineteen *sox* genes are also present in singleton in the three other Osteichthyes : the coelacanth, the mouse and the human. Of note, the elephant shark, a cartilaginous fish, seems to have lost *sox19*, as no sequences corresponding to this gene have been identified in its genome, using all the detection methods available. Considering teleosts, this analysis indicates an enrichment of the Sox family. Indeed, 57,8% (11/19) of vertebrate *sox* genes are duplicated in teleostean fishes. Over these duplicates, 18,2% (2/11) are duplicated in all the studied species, whereas 81,1% (9/11) are duplicated only in some lineages and or species. Finally, 36,8% (7/19) *sox* genes are in singleton in all the species analyzed. Interestingly, all the teleostean species does not exhibit the same *sox* gene content, reflecting that, during the evolution, different events of lineage-specific loss and retention occurred. Moreover, no corresponding sequences have been identified for the three vertebrate

sox10, *sox12* and *sox30* respectively in the cavefish genome, in the acanthopterygian genomes, and in six teleost genomes.

Origin of sox gene duplicates in teleosts

As described previously (figure 1), the Sox family in teleostean fishes is characterized by the presence of a lot of duplicates. However, duplicated genes can arise through small-scale duplication (SSD) or whole genome duplication (WGD). In this study, we determined by phylogeny and synteny analyses, that, except for one specific case (*sox17*), all *sox* duplicates still present in teleostean genomes are coming from the teleost-specific WGD (figure 2 and supplementary figure 1, A2-3 to K2-3). Indeed, as shown on figure 2, duplicated *sox* genes are localized on ohnologue chromosomes (paralogue chromosomes derived from WGD; Woods *et al.*, 2005; Kasahara *et al.*, 2007) for the four species zebrafish, medaka, stickleback and tetraodon. The so-called *sox32* is a duplicate of *sox17* arise through SSD, and its identification in all teleost genomes suggests that this event took place in the common ancestor of teleost. Of note, the salmon *Salmo salar* possesses more *sox* genes than the other teleostean species presented in this study due to the specific WGD that occurred in the salmonidae about 60 mya (supplementary table 1, Crête-Lafrenière *et al.* 2012). To conclude, in teleost genomes, 11 *sox* genes are still in duplicates, of which 10 are derived from the teleost-specific WGD. As mentioned above, retain of both copies suggest evolutionary mechanisms that counteract deletion event and/or pseudogenization. Sub-functionalization and/or neo-functionalization, that involve gene expression retention and functional divergence, are classic models to explain duplicate gene retention. We thus investigated molecular evolution changes after duplication to test the applicability of these explanations for the retention of the five *sox4*, *sox8*, *sox9*, *sox10* and *sox11*, which are still in duplicates in the majority of studied teleostean species. As a reminder, the two *sox4* and *sox11* belong to the SoxC group, whereas the three *sox8*, *sox9* and *sox10* correspond all together to the SoxE group in vertebrates.

Molecular evolution of the protein-coding and non-coding sequences

The two sub-functionalization and neo-functionalization models make different predictions about the rates of protein-coding and non-coding sequences evolution of the duplicates (Chain & Evans 2006, Innan & Kondrashov 2010). We performed two analyses used by Chain & Evans (2006) to test specific predictions about the evolutionary rates (in terms of dN/dS ratio) in the protein-coding sequences after

duplication (table 1). The first analysis tests a subfunctionalization model (Model B) that has a higher dN/dS ratio after duplication (and thus relaxed purifying selection and/or adaptive evolution on both paralogues after duplication). This was compared to a null model with no change in the dN/dS ratio after duplication (Model A, in which a single dN/dS ratio is estimated for all branches). With this analysis, we show that *sox9* paralogues have a significantly higher average dN/dS ratio than their non-teleost orthologues, and we therefore inferred relaxed purifying selection on both paralogues after duplication. The second analysis tests a neofunctionalization model (Model C) that has a different dN/dS ratio on each paralogue. This was compared to the Model B, in which dN/dS ratio are equal in each paralogue. The likelihood of the Model C was significantly higher than the null model for the other *sox* duplicates. In these cases, compared to non-teleost orthologues, one paralogue evolved significantly faster than its paralogue, with an increase of the dN/dS ratio, suggesting relaxation of purifying selection and/or adaptive evolution in the faster paralogue. This may indicate neofunctionalization in the accelerated copy. Finally, for all pairs, we found a dN/dS $\ll 1$ (and close to 0), indicating that purifying selection after duplication is pervasive. For example, the average dN/dS ratio of the *sox4* is estimated closed to 0,03. This is in agreement with the crucial role in development and other key biological processes of Sox proteins, which are likely to be under strict evolutionary constraints.

For each duplicates, we then investigated the conservation of conserved non-coding elements (CNEs) within ray-finned fishes and within vertebrates (using the two tetrapods human and mouse). For this analysis, we used the spotted-gar as a reference (figure 3, a to e). The number of CNEs varies widely according to the *sox* genes : few CNEs are detected in the neighboring of *sox8* and *sox10* (six and three respectively), whereas several are identified in the vicinity of *sox4*, and even more when looking for *sox11* and *sox9*. For instance, we identified 37 CNEs around *sox9*, among which the vast majority are ray-finned fish-specific. Indeed, only 20% of these CNEs are shared by tetrapods (human and mouse in this study) and ray-finned fish. Moreover, except for *sox9*, the CNEs identified tend to be asymmetrically distributed between the two paralogues in teleosts. For example, 87% of the 36 CNEs detected in the vicinity of *sox11* are kept on the single *sox11a* in teleosts. On the contrary, CNEs shared by tetrapods and ray-finned fish tend to be distributed symmetrically between paralogues. To note, conserved non-coding elements are supposed to be functional elements, such as *cis*-regulatory elements (Hemberg *et al.* 2012). Thus, this study show that *sox4*, *sox8*, *sox10* and *sox11* paralogues evolve asymmetrically in their protein-coding sequences and in

their potential *cis*-regulatory elements, in agreement with neofunctionalization scenario in the faster paralogue. Moreover, the differential loss of certain potential *cis*-regulatory elements between the two teleost *sox9* genes supports the subfunctionalization scenario inferred above.

Expression pattern during embryogenesis of sox duplicates

Still in order to understand the mechanisms leading to duplicates retention after, in our study, an event of WGD, we then investigated the expression pattern of the five duplicated *sox* genes (4, 8, 9, 10 and 11) during embryogenesis. The two classic fishes, the zebrafish *Danio rerio* (Dr) and the medaka *Oryzias latipes* (Ol), have been used. We performed qRT-PCR experiments on eight developmental stages representing five crucial periods of embryonic development: segmentation (S), gastrula (G), neurula (N), pharyngula (P) and hatching (H) (from Kimmel et al., 1995 and Iwamatsu et al., 2004). The results obtained are presented figure 4A. Two levels of interpretation are possible: an intra-species level (*i.e.* we compare the expression pattern of two duplicates within one species) and an inter-species level (*i.e.* we compare the expression profile of two genes between the two species used). The intra-species analysis gives an idea on the evolution of two paralogues after an event of WGD and can allow understanding why both are retained. On the other hand, the inter-species analysis is suited to assess questions dealing with differential evolution, genomic incompatibility, population isolation and speciation if we consider the ultimate situation.

Considering the first level of analysis, 4 genes are still in duplicates in the zebrafish genome: *sox4*, *sox8*, *sox9* and *sox11* (fig1). No strong differences of expression are detected between *sox4a,b*; *sox9a,b* and *sox11a,b* during the zebrafish embryogenesis (fig4A). However, we noticed that the maximum level of expression is different between *sox9a* and *b*: *sox9a* is highly expressed during the pharyngula and the hatching stages, while *sox9b* is highly expressed during the neurula stage. Interestingly, *sox8* paralogues show a completely different expression pattern: *sox8b* is mainly expressed at the early stage of embryogenesis with a decreasing level of expression until the end of gastrula, while absolutely no expression is detected for *sox8a*. This result suggests a differential evolution of the expression pattern of *sox8* duplicates in the zebrafish ancestor, which can probably explain why both paralogues are still present in the genome of this fish. In the medaka genome, three genes are duplicated: *sox4*, *sox9* and *sox10*. *Sox4a,b* and *sox10a,b* present similar expression profiles during medaka development (fig4A). In contrast, *sox9a* is mainly expressed during segmentation and hatching periods, whereas

sox9b expression is spread from the gastrula to the hatching, suggesting, as for *sox8* above, a differential expression pattern evolution of *sox9* paralogues in the medaka.

At the inter-species level, *sox8*, *sox10* and *sox11* are not duplicated in both species: the zebrafish genome possesses two copies of *sox8* and *sox11*, while *sox10* is duplicated in the medaka genome (fig1). For *sox10* and *sox11*, the expression pattern of the two paralogues in one species is similar of the corresponding singleton in the other species. The situation of *sox8* is particular as first *sox8a* is not expressed during zebrafish development (fig4A), and second the expression profiles of *sox8b* in both species are completely different: *sox8b* is expressed from segmentation to gastrula stages in the zebrafish embryo, while the medaka *sox8b* is expressed during segmentation and then highly during pharyngula and hatching periods. On the contrary, *sox4* and *sox9* are both duplicated in the two teleostean genomes. As shown figure 4A, if the expression profiles of each orthologues are overlapping, lineage-specific expression patterns are also detected, as for example *sox9a* which is highly expressed during the segmentation in the medaka embryo only.

All together, these results suggest a lineage-specific evolution of the expression pattern of paralogues and/or orthologues.

Expression pattern during adulthood of sox duplicates

Three fish species have been used in this study (the zebrafish, the medaka and the platyfish *Xiphophorus maculatus* Xm). The mouse has been also introduced for comparison as a tetrapod. Seven males and females adult tissues have been collected and analysed: ovaries, testes, spleen, liver, muscles, eyes and brain. Results obtained are presented figure 4B. After analysis, three categories of duplicates/paralogues have been identified: 1) the expression of the paralogues in the three fish species analysed, as in the mouse, is equivalent in term of location, even if we can observe minor variations in term of level of expression. We place in this category medaka and platyfish *sox10* paralogues, medaka, platyfish and zebrafish *sox4* paralogues (even if we noticed the fact that the zebrafish *sox4a* copy has lost the expression in gonads), and finally medaka, platyfish and zebrafish *sox9* paralogues. Considering *sox9* paralogues, the expression of both copies *a* and *b* recapitulates the expression observed in mouse tissues, leading to the conclusion that the current observed expression of *sox9* genes in the studied species might reflect the expression of the ancestral *sox9* gene present in the genome of the common ancestor of teleost and tetrapods. However, it seems that in the platyfish we detect a specific distribution of duplicates expression leading to a specialization of

paralogues as *sox9a* is highly expressed in ovaries while *sox9b* is mainly expressed in the liver. This situation could represent a case of sub-functionalization *via* the progressive specialization of the expression of both paralogues in specific tissues. 2) the supposed “ancestral” expression (reflected by the expression observed in the mouse) of the gene at the origin of the obtained paralogues, i.e. the profile of expression of the gene present in the common ancestor of teleostean and tetrapods before the split and before the fish-specific WGD, is distributed between the two paralogues. This is the case for *sox8* genes in the platyfish and the zebrafish. This kind of repartition that we can assimilate to a specialization is also accompanied by a huge increase of the tissue-specific expression. 3) paralogues show the same expression profile as the supposed ancestral gene in comparison with what we observe in the mouse for the corresponding orthologue, but they in addition present new specific expression sites. This is the case for *sox11*. Indeed, in the mouse, *sox11* is expressed only in the brain. However, in the zebrafish and the platyfish, *sox11a* is expressed in the brain mainly, but also in the eyes and the ovaries considering the latest species, while *sox11b* is expressed at a little level in the brain, but highly in the ovaries for both species. Two hypothesis can explain such a situation: i) the ancestral *sox11* gene had an expression restricted to the brain as it is observed in the mouse, and teleostean, thanks to the fish-specific WGD, acquired new sites of expression (eyes and ovaries) reflecting a case of neo-functionalization; ii) the ancestral *sox11* gene had the current observed fish expression pattern, and tetrapods have lost eyes and ovaries sites of expression.

Discussion

Expansion of the Sox family during the teleost-specific whole genome duplication

This study of comparative genomics in the vertebrate lineage shows the great expansion of the Sox transcription factors family in teleostean fishes. Furthermore, we demonstrate by syntenic and phylogenetic analyses that this expansion is the consequence of the well-established teleost-specific whole genome duplication estimated around 226-316 mya (Jaillon *et al.* 2004, Kasahara *et al.* 2007, Van de Peer *et al.* 2009). Four pairs of *sox* genes, *sox9*, *sox10*, *sox11* and *sox19*, were previously suggested by previous phylogenetic reconstructions and/or microsynteny to be derived from the teleost-specific WGD (Chiang *et al.* 2001, Hoegg *et al.* 2004, Okuda *et al.* 2006, Braasch *et al.* 2009) and 6 pairs are clearly identified as it in this study. As mentioned

above, the observed retention rate of duplicates of 55,5% (10/18) after the WGD for teleost *sox* genes is higher than the already described average deduced from genome-wide analyses in teleosts (between 12% and 24%, Postlethwait *et al.* 2000, Jaillon *et al.* 2004, Woods *et al.* 2005, Brunet *et al.* 2006, Kassahn *et al.* 2009). This observation confirms the preferential retention and amplification of a number of gene families over successive WGDs as transcription factors (Blanc & Wolfe 2004, Brunet *et al.* 2006, Aury *et al.* 2006, Kassahn *et al.* 2009). Accordingly, the retention rate of duplicates in the salmon *Salmo salar* for *sox* genes after the additional and more recent WGD that occurred in salmonidae (around 50 to 100 mya) is higher than 80% (Supplementary data). This result is in agreement with the hypothesis suggesting that salmonids genomes are still in the re-diploidization phase and consequently still present a huge number of duplicates. Finally, this study indicates that *sox* paralogues obtained by single-gene duplication (SGD) are rare in the teleost lineage. Indeed, only one case have been characterized: the two *sox17* and *sox32* detected in all analysed teleost genomes are paralogues but, contrary to the other teleost *sox* duplicates, not the consequence of the teleost-specific WGD. Thus, we identified only one event of *sox*-SGD that certainly occurred in the common ancestor of teleosts. Consequently, these observations suggest that the Sox transcription factors that expand mainly by WGD rather than SGD are dosage-balanced genes sensitive for which changes in the relative quantity of one gene product only should be deleterious (Papp *et al.* 2003). Accordingly, the expansion of the Sox family in stem vertebrates, as many other transcription factors families, seems mainly due to the two whole genome duplications that occurred at the base of the vertebrate lineage (Hokamp *et al.* 2003, Okuda *et al.* 2006, Holland 2009). Interestingly, some authors argue that WGD events might allow genetic innovations and diversification that would have been unlikely to arise otherwise (Freeling & Thomas 2006, Conant & Wolfe 2008). This idea implies that, in particular, the expansion of the Sox family during the teleost-specific WGD, and the subsequent high retention rate of duplicates (providing raw material for evolutionary novelties), have certainly been a key event in the teleosts diversification considering their essential role in morphology, physiology and behavior of vertebrates (Sarkar & Hochedlinger 2013).

sox genes paralogues tend to evolve asymmetrically after the WGD

We also used a comparative genomics approach to study the molecular evolution of five *sox* duplicates at different levels of genomic organization. This study shows that *sox* duplicates tend to evolve asymmetrically in both their coding sequence and potential

cis-regulatory elements, and that this evolution is lineage-specific. Moreover, microsyntenic analyses also suggest asymmetric and species-specific rates of chromosomal rearrangement between paralogs. Accordingly, such asymmetric evolution after the teleost-specific WGD, and more generally after WGD, is already reported in several studies (Wagner *et al.* 2005, Braasch *et al.* 2006, Brunet *et al.* 2006, Steinke *et al.* 2006, Fares *et al.* 2006). Of note, the relative contribution of both relaxed negative selection and positive selection in this asymmetric evolution is rather difficult to estimate due to the ancientness of the teleost-specific WGD and the motley nature of paralogue evolution (between lineage, genomic locations of each gene copy, saturation of synonymous substitution) over several hundreds millions of years (Raes & Van de Peer 2003, Chain & Evans 2006, Steinke *et al.* 2006, Conant & Wolfe 2008). Nonetheless, comparative genomics analyses with non-teleost species shows that one copy tend to be more constraint in its coding sequence as well as in its genomic location and potential *cis*-regulatory elements than the rapidly evolving copy. This observation suggests concerted strict evolutionary constraints both in biochemical properties of the corresponding protein and its regulation of expression. On the contrary, the second copy seems to freely evolve, suggesting relaxing constraints in the whole paralogon. As well, this observation suggests that this copy can “develop” new biochemical innovation and/or new site of expression (by degenerative mutations or adaptive evolution). However, *sox* duplicates phylogenies, and their respective retention patterns of regulatory elements suggest that the evolution of the divergent paralogue is lineage-specific. Indeed, this evolution is not characterized by a unique event of rapid evolution of one copy just after the whole genome duplication. It seems that there is a lineage-specific accumulation of mutations (adaptive or degenerative) in the coding and non-coding sequences, leading to potential independent neo-sub-functionalization or pseudogenization (the divergent paralogue is lost from some teleost genomes). Thus, this study suggests that *sox* duplicates produced by the teleost-specific WGD tend to evolve asymmetrically: one copy evolving under strict evolutionary constraints and has conserved most of the protein features and potential regulatory elements compared to the non-teleost orthologues (and thus the hypothetic ancestral gene); the second copy evolving through released selection and has diverged in term of coding and non-coding sequences in a lineage-specific manner.

Lineage-specific evolution of sox duplicates and consequences

This study of functional comparative genomics between the zebrafish, the medaka and the platyfish clearly indicates that orthologues in the different species have developed different expression patterns. For example, species-specific differences in expression are detected for *sox8a* between adult zebrafish and platyfish, and for *sox9a* and *sox9b* between adult zebrafish and medaka. Of note, these results are in agreement with previous studies that infer species-specific sub-functionalization for *sox9* duplicates in teleosts (Cresko et al. 2003, Kl ver et al. 2005). Interestingly, the medaka and the platyfish, that are more closely related species, also show specific-species differences in *sox* paralogues expression pattern. For instance, the platyfish *sox9b* is expressed in liver, eyes and brain while its orthologue in the medaka is expressed in testes, eyes and brain. These observations certainly are the consequences of lineage-specific sub-function partitioning and/or neo-functionalization after the teleost-specific WGD. In addition, this study highlights recurrent non-functionalization events of the same paralogue in different species. Indeed, *sox14b* is lost from the three medaka, tetraodon and stickleback genomes, and *sox11b* from the three medaka, tetraodon and takifugu genomes. This suggests less functional constraints in this paralogue, in which accumulation of mutations should have led by chance to non-functionalization in some species, or to sub-functionalization and/or neo-functionalization in others species. The teleost-specific high expression of *sox11b* in ovaries could be an example of neo-functionalization in zebrafish and platyfish, but this hypothesis is not the only one that could explain the observed expression pattern. In agreement, the increase of the dN/dS ratio in the corresponding coding-sequences indicates relaxed purifying selection in the paralogue recurrently lost in different genomes (table 1 and supplementary table 2). The differential sub-functionalization and/or neo-functionalization of *sox* duplicates after the teleost-specific WGD, as well as species-specific non-functionalization, could have facilitated speciation by a Dobzhansky-Muller process (Lynch & Force 2000). Moreover, this study suggests that these events sometimes have occurred tens of millions of years after the teleost-specific-WGD, such as in medaka and platyfish, and supports the hypothesis that this genetic mechanisms can contribute to speciation over a long timeframe after WGD (Sémon & Wolfe 2007, Van de Peer et al. 2009). Finally, reciprocal gene loss is also a well-established genetic mechanism that can promote speciation after whole genome duplication (Lynch & Force 2000, Volff 2005, Scannell et al. 2006, Van de Peer et al. 2009). However, the macrosynteny analyses of *sox* genes that are in singleton in all teleosts show that they are all orthologues and localized on

orthologue chromosomes (supplementary figure 2). This indicates that their ancestral loci do not undergone reciprocal gene loss.

References

- Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y., et al. (1998). Zebrafish hox clusters and vertebrate genome evolution. *SCIENCE* 282, 1711–1714.
- Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552.
- Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444, 171–178.
- Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678.
- Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.
- Bowles, J., Schepers, G., and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol.* 227, 239–255.
- Braasch, I., Salzburger, W., and Meyer, A. (2006). Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogs involved in teleost coloration. *Mol. Biol. Evol.* 23, 1192–1202.
- Braasch, I., Brunet, F., Volff, J., and Schartl, M. (2009). Pigmentation Pathway Evolution after Whole-Genome Duplication in Fish. *GENOME Biol. Evol.* 1, 479–493.
- Brunet, F., Crollius, H., Paris, M., Aury, J., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* 23, 1808–1816.
- Chain, F.J.J., and Evans, B.J. (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genet.* 2, e56.
- Chiang, E.F.-L., Yan, Y.-L., Guiguen, Y., Postlethwait, J., and Chung, B. (2001). Two Cyp19 (P450 Aromatase) Genes on Duplicated Zebrafish Chromosomes Are Expressed in Ovary or Brain. *Mol. Biol. Evol.* 18, 542–550.
- Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* 9, 938–950.
- Cresko, W., Yan, Y., Baltrus, D., Amores, A., Singer, A., Rodriguez-Mari, A., and Postlethwait, J. (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev. Dyn.* 228, 480–489.
- Crête-Lafrenière, A., Weir, L.K., and Bernatchez, L. (2012). Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PloS One* 7, e46662.
- Dehal, P., and Boore, J. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLOS Biol.* 3, 1700–1708.
- Fares, M.A., Byrne, K.P., and Wolfe, K.H. (2006). Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. *Mol. Biol. Evol.* 23, 245–253.

- Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16, 805–814.
- Galtier, N., Gouy, M., and Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* CABIOS 12, 543–548.
- Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* 346, 245–250.
- Haldane JBS. 1933. The part played by recurrent mutation in evolution. *Am Nat* 67: 5–9.
- Hawkins, J.R. (1993). Mutational analysis of SRY in XY females. *Hum. Mutat.* 2, 347–350.
- Hoegg, S., Brinkmann, H., Taylor, J., and Meyer, A. (2004). Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59, 190–203.
- Hokamp, K., McLysaght, A., and Wolfe, K.H. (2003). The 2R hypothesis and the human genome sequence. *J. Struct. Funct. Genomics* 3, 95–110.
- Holland, L. (2009). Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nat. Rev. Neurosci.* 10, 736–746.
- Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108.
- Iwamatsu, T. (2004). Stages of normal development in the medaka *Oryzias latipes*. *Mech. Dev.* 121, 605–618.
- Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *NATURE* 431, 946–957.
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *GENOME Res.* 20, 1313–1326.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719.
- Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C., and Ragan, M.A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19, 1404–1418.
- Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624.
- Kimmel, C., BALLARD, W., KIMMEL, S., ULLMANN, B., and SCHILLING, T. (1995). STAGES OF EMBRYONIC-DEVELOPMENT OF THE ZEBRAFISH. *Dev. Dyn.* 203, 253–310.
- King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., et al. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451, 783–788.
- Kluver, N., Kondo, M., Herpin, A., Mitani, H., and Schartl, M. (2005). Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. *Dev. GENES Evol.* 215, 297–305.

- Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods San Diego Calif* 25, 402–408.
- Lynch, M., and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *SCIENCE* 290, 1151–1155.
- De Martino, S., Yan, Y., Jowett, T., Postlethwait, J., Varga, Z., Ashworth, A., and Austin, C. (2000). Expression of sox11 gene duplicates in zebrafish suggests the reciprocal loss of ancestral gene expression patterns in development. *Dev. Dyn.* 217, 279–292.
- Mavropoulos, A., Devos, N., Biemar, F., Zecchin, E., Argenton, F., Edlund, H., Motte, P., Martial, J., and Peers, B. (2005). sox4b is a key player of pancreatic alpha cell differentiation in zebrafish. *Dev. Biol.* 285, 211–223.
- Nüsslein-Vollhard C, Dahm R: Zebrafish: a practical approach Oxford University Press, USA; 2002.
- Ohno S. 1970. Evolution by gene duplication. Springer Verlag, Berlin.
- Okuda, Y., Yoda, H., Uchikawa, M., Furutani-Seiki, M., Takeda, H., Kondoh, H., and Kamachi, Y. (2006). Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. *Dev. Dyn.* 235, 811–825.
- Otto, S.P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462.
- Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197.
- Van de Peer, Y., Maere, S., and Meyer, A. (2009). OPINION The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732.
- Phochanukul, N., and Russell, S. (2010). No backbone but lots of Sox: Invertebrate Sox genes. *Int. J. Biochem. CELL Biol.* 42, 453–464.
- Postlethwait, J., Amores, A., Cresko, W., Singer, A., and Yan, Y.-L. (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet.* 20, 481–490.
- Raes, J., and Van de Peer, Y. (2003). Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl. Bioinformatics* 2, 91–101.
- Santos, M.E., and Salzburger, W. (2012). How Cichlids Diversify. *Science* 338, 619–621.
- Sarkar, A., and Hochedlinger, K. (2013). The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell* 12, 15–30.
- Semon, M., and Wolfe, K. (2007). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512.
- Sémon, M., and Wolfe, K.H. (2007). Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet. TIG* 23, 108–112.
- Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell-Badge, R., and Goodfellow, P.N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* 346, 240–244.
- Steinke, D., Salzburger, W., Braasch, I., and Meyer, A. (2006). Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* 7, 20.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24, 1596–1599.

Volff, J. (2005). Genome evolution and biodiversity in teleost fish. *HEREDITY* 94, 280–294.

Wagner, G.P., Takahashi, K., Lynch, V., Prohaska, S.J., Fried, C., Stadler, P.F., and Amemiya, C. (2005). Molecular evolution of duplicated ray finned fish HoxA clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. *J. Mol. Evol.* 60, 665–676.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.

Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. (2005). The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* 15, 1307–1314.

Zhang, L., Lin, D., Zhang, Y., Ma, G., and Zhang, W. (2008). A homologue of Sox11 predominantly expressed in the ovary of the orange-spotted grouper *Epinephelus coioides*. *Comp. Biochem. Physiol. B-Biochem. Mol.* 149, 345–353.

Zhang, L., Zhu, T., Lin, D., Zhang, Y., and Zhang, W. (2010). A second form of Sox11 homologue identified in the orange-spotted grouper *Epinephelus coioides*: Analysis of sequence and mRNA expression patterns. *Comp. Biochem. Physiol. B-Biochem. Mol.* 157, 415–422.

Figure captions

Figure 1: Landscape of sox duplicates present in fifteen vertebrate genomes.

The 19 vertebrate *sox* genes (on top) are divided into the 7 groups B1, B2, C, D, E, F and H. Phylogenetic relationships of the different analyzed vertebrate species are shown on the left. The orange circle on the phylogeny represents the teleost-specific WGD. The analyses were performed on the five non-teleost species elephant shark *Callorhynchus milii* (Cm), mouse *Mus musculus* (Mm), human *Homo sapiens* (Hs), coelacanth *Latimeria chalumnae* (Lc) and spotted-gar *Lepisosteus oculatus* (Lo) and the ten teleost species cave fish *Astyanax mexicanus* (Am), zebrafish *Danio rerio* (Dr), salmon *Salmo salar* (Ss), cod *Gadus morhua* (Gm), tilapia *Oreochromis niloticus* (On), medaka *Oryzias latipes* (Ol), platyfish *Xiphophorus maculatus* (Xm), tetraodon *Tetraodon nigroviridis* (Tn), fugu *Takifugu rubripes* (Tr) and stickleback *Gasterosteus aculeatus* (Ga). Light blue and orange squares show respectively singleton and duplicates produced by WGD. Dark blue squares show duplicates produced by small-scale duplication (SSD). Grey squares indicate non-detected copy in the genome. The mammals-specific SoxA group is not shown on the figure.

Figure 2: Macrosynteny of the teleost-specific WGD-duplicated sox genes.

Grey lines connect paralogue genes on the different chromosomes or groups in the zebrafish (a), medaka (b), stickleback (c) and tetraodon (d) genomes (Chromosome or group numbers are indicated in grey). Orange lines connect paralogue *sox* genes on chromosomes. In the case of neighboring genes *sox8* and *sox9*, a single curve represents the link between the two duplicates.

Figure 3: Non-coding and coding sequences evolution of five sox duplicates.

Conserved non-coding elements (CNEs) are highlighted in the vicinity of fish *sox* genes, either present up to human or being fish-specific, with the spotted-gar as a reference. Blue and orange circles represent CNEs respectively in the vicinity of paralogue a and paralogue b of *sox4* (a), *sox11* (b), *sox8* (c), *sox9* (d) and *sox10* (e).

Figure 4: Expression pattern of five sox duplicates in three teleost species.

(a) Expression of five duplicated *sox* gene during the *Danio rerio* (*Dr*) and *Oryzias latipes* (*Ol*) development. The five major developmental periods are defined on the left. (b) Expression of five duplicated *sox* gene in seven adult tissues of *Danio rerio* (*Dr*), *Oryzias latipes* (*Ol*), *Xiphophorus maculatus* (*Xm*) and *Mus musculus* (*Mm*). *Sox* number are shown at the bottom and paralogues a and b are presented on top. Blue and grey squares indicate gene expression and no expression, respectively. The intensity of the blue is defined from the $\Delta\Delta C_t$ value (see material & methods) and darkens with expression increase.

Table 1: Comparison of alternative models with different branch-specific dN/dS ratio in each paralogue for 5 pairs of *sox*.

Supplementary figure 1. Phylogenetic reconstructions and macrosynteny analyses of *sox1* (a1-3), *sox4* (b1-3), *sox6* (c1-3), *sox8* (d1-3), *sox9* (e1-3), *sox10* (f1-3), *sox11* (g1-2), *sox14* (h1-3), *sox21* (i1-2), *sox17/32* (j1-3) and *sox19* (k) in teleostean fishes. Parologue a and parologue b are respectively highlighted in blue and orange.

Supplementary figure 2. Macrosynteny analyses and orthology relationships of *sox2* (a), *sox3* (b), *sox5* (c), *sox13* (d), *sox7* (e) and *sox18* (f) in the four zebrafish *Danio rerio* (*Dr*), medaka *Oryzias latipes* (*Ol*), tetraodon *Tetraodon nigroviridis* (*Tn*), and stickleback *Gasterosteus aculeatus* (*Ga*).

Supplementary table 1. Accession numbers or scaffold references of the sequences used in this study.

Supplementary table 2. Comparison of alternative models with different branch-specific dN/dS ratio in each paralogue for 9 pairs of *sox*.

Figure 1

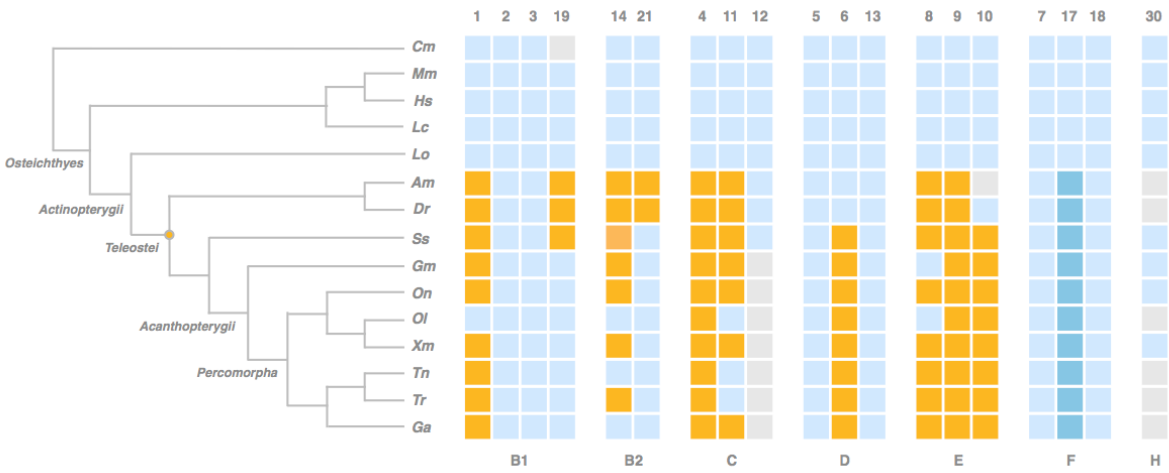


Figure 2

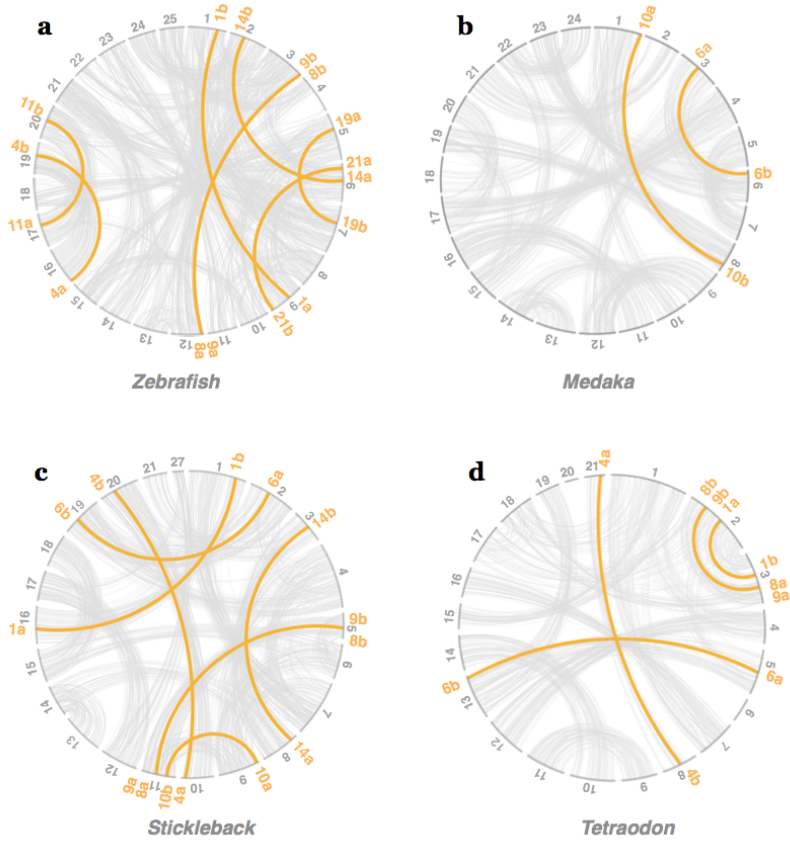


Figure 3

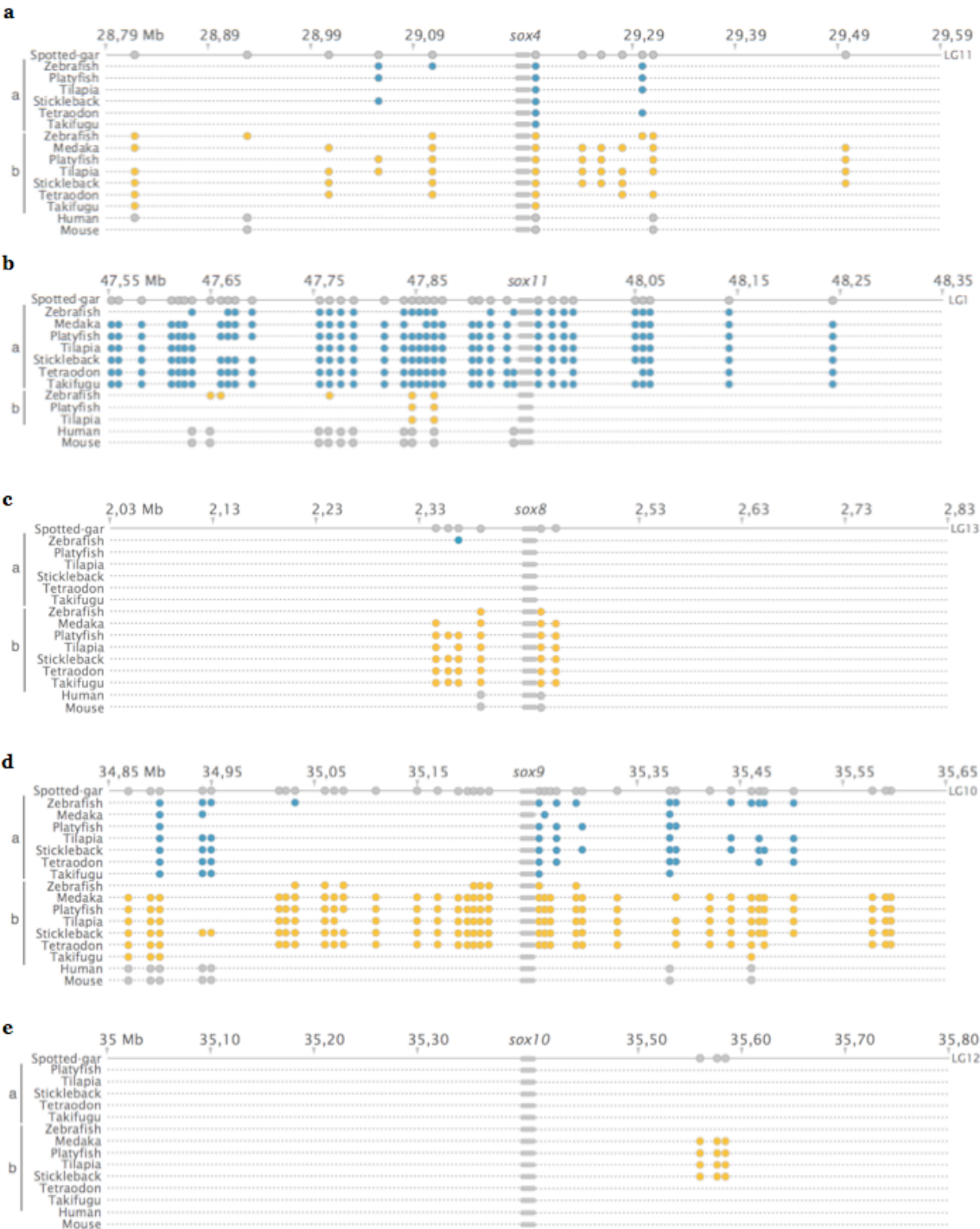


Figure 4

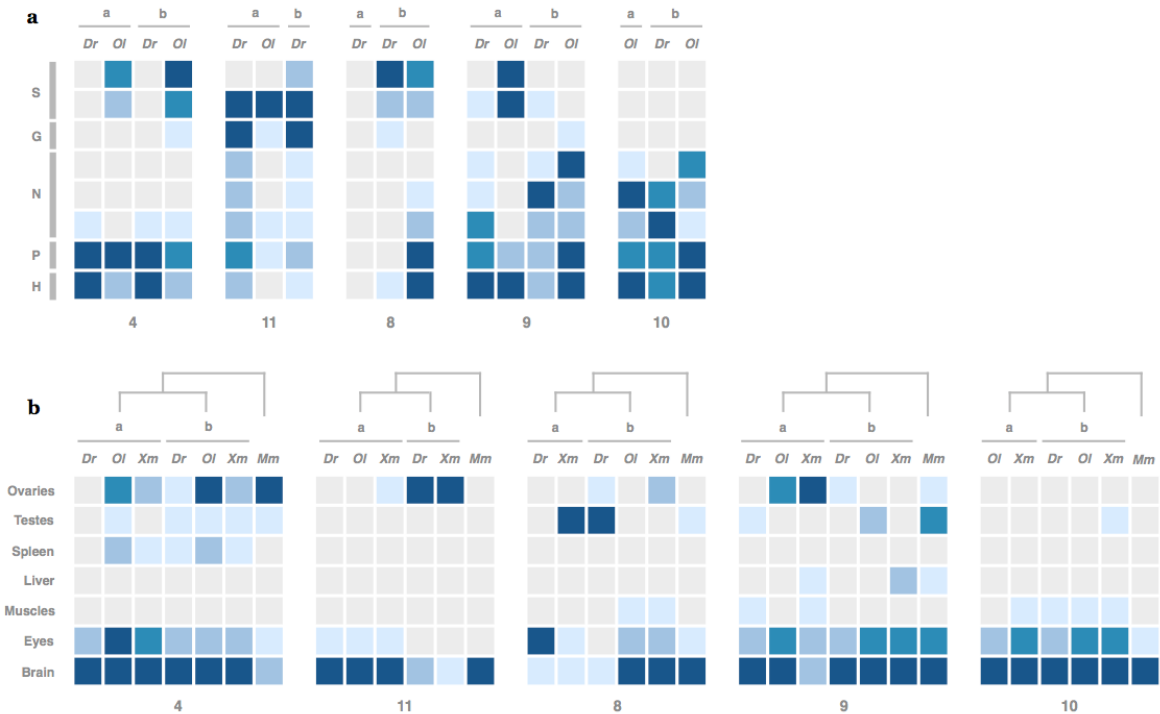
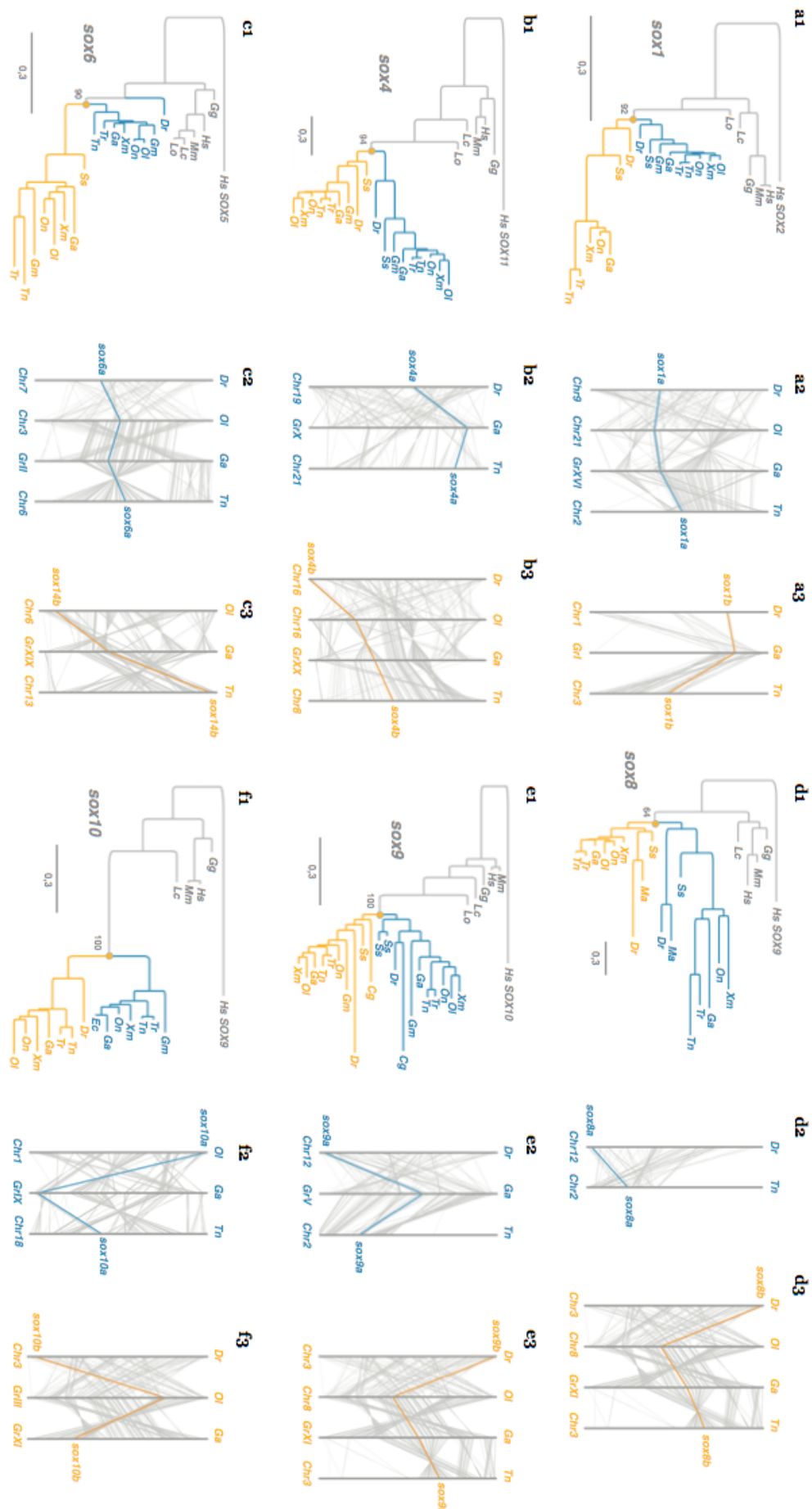
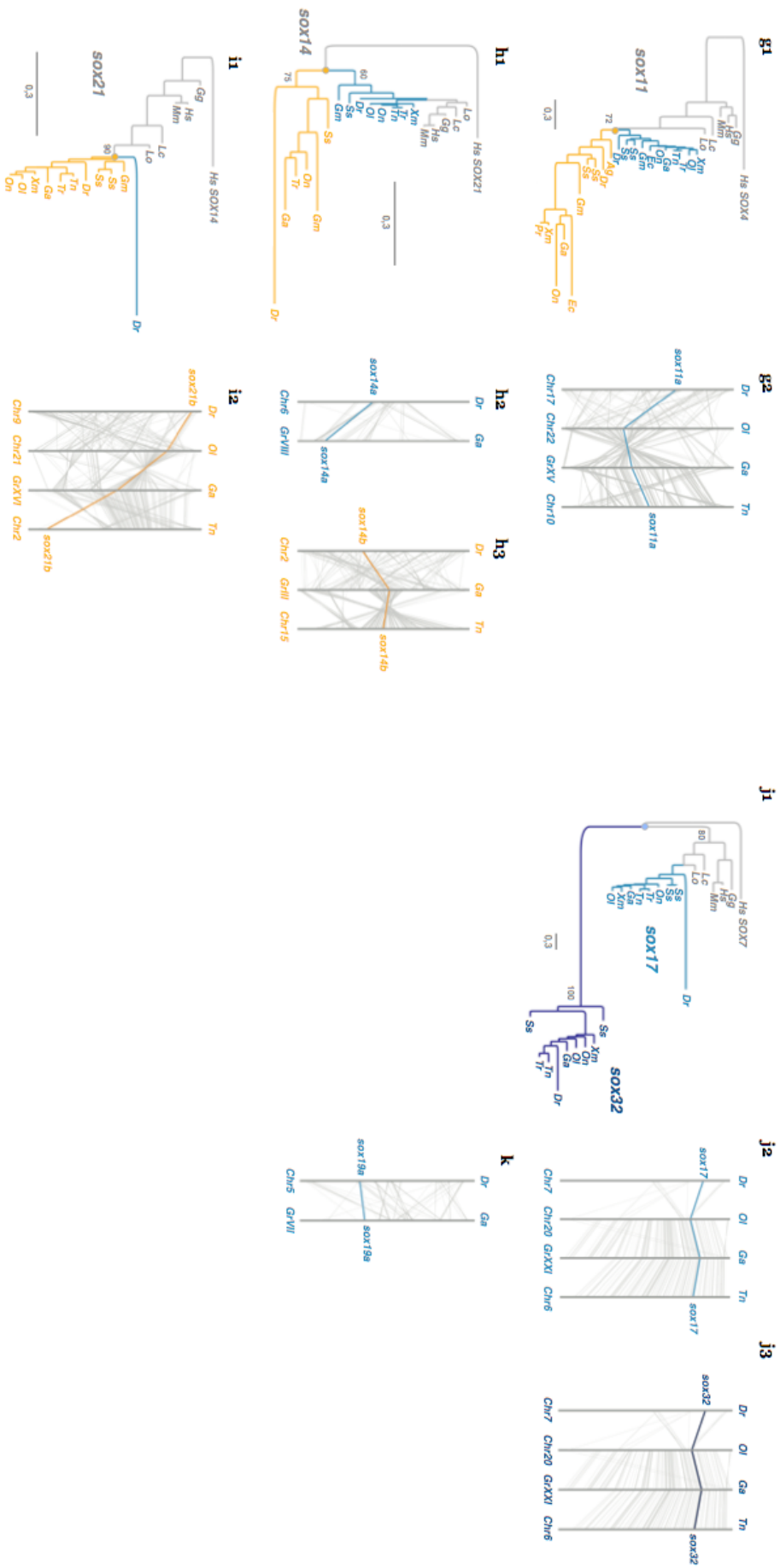


Table 1

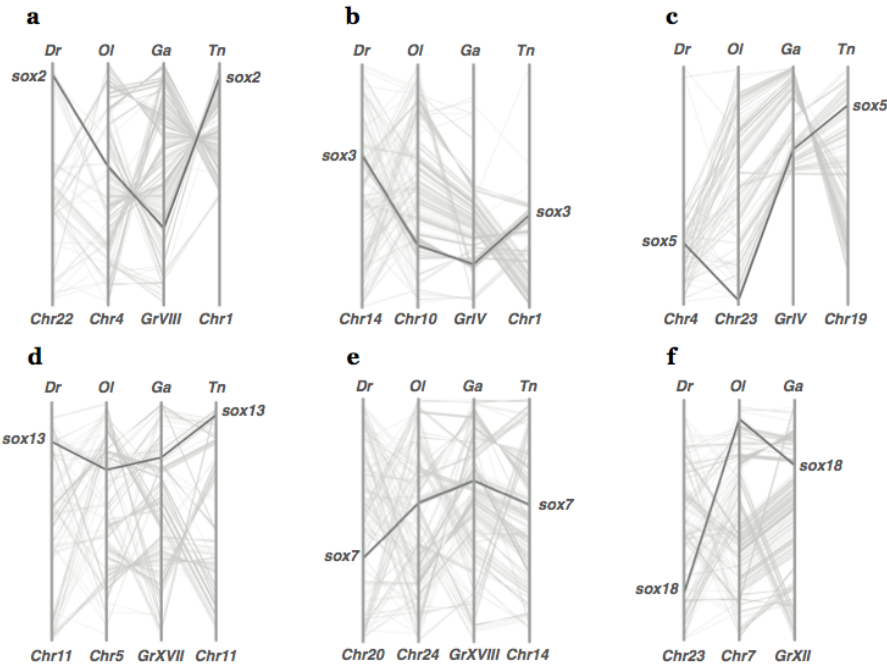
sox	Model	LnL (np)	Estimates of ω			Tests	df	p-value	Putative mechanism for retained expression (Chain & Evans 2006)
4	A : One ratio	-7271.80	0,032			Model A vs. B	1	NS	
	B : two ratio	-7270.88	0.028	0.034		Model B vs. C	1	< 0,05	Neofunctionalization
	C : tree ratio	-7268.92	0.029	0.029	0.040				
11	A : One ratio	-8648.14	0,051			Model A vs. B	1	< 0,05	
	B : two ratio	-8643,88	0,039	0,057		Model B vs. C	1	<0,0001	Neofunctionalization
	C : tree ratio	-8631,77	0,039	0,038	0,077				
8	A : One ratio	-8608.62	0,044			Model A vs. B	1	<0,0001	
	B : two ratio	-8594,53	0,021	0,056		Model B vs. C	1	<0,0001	Neofunctionalization
	C : tree ratio	-8552,25	0,021	0,026	0,10				
9	A : One ratio	-11100.46	0,029			Model A vs. B	1	<0,001	
	B : two ratio	-11094,96	0,020	0,032		Model B vs. C	1	NS	Relaxed purifying selection in both paralogues
	C : tree ratio	-11095,11	0,020	0,033	0,031				
10	A : One ratio	-9970.81	0,051			Model A vs. B	1	< 0,05	
	B : two ratio	-9968,20	0,040	0,056		Model B vs. C	1	< 0,05	Neofunctionalization
	C : tree ratio	-9964,26	0,040	0,044	0,066				

Supplementary Figure 1





Supplementary Figure 2



Supplementary table 2

sox	Model	LnL (np)	Estimates of ω			Tests	df	p-value	Putative mechanism for retained expression (Chain & Evans 2006)
			ω^r	ω^p / ω^a	ω^b				
B	1	A : One ratio	-6811,052	0,028		Model A vs. B	1	NS	
		B : two ratio	-6809,73	0,022	0,030	Model B vs. C	1	<0,0001	Neofunctionalization
		C : tree ratio	-6801,21	0,022	0,020	0,042			
	14	A : One ratio	-5843,9	0,036		Model A vs. B	1	<0,0001	
		B : two ratio	-5829,83	0,017	0,049	Model B vs. C	1	<0,0001	Neofunctionalization
		C : tree ratio	-5797,891	0,017	0,018	0,091			
	21	A : One ratio	-5821,28	0,020		Model A vs. B	1	NS	
		B : two ratio	-5819,52	0,014	0,023	Model B vs. C	1	<0,0001	Neofunctionalization
		C : tree ratio	-5811,101	0,013	0,073	0,021			
C	4	A : One ratio	-7271,80	0,032		Model A vs. B	1	NS	
		B : two ratio	-7270,88	0,028	0,034	Model B vs. C	1	<0,05	Neofunctionalization
		C : tree ratio	-7268,92	0,029	0,040	0,029			
	11	A : One ratio	-8648,14	0,051		Model A vs. B	1	<0,05	
		B : two ratio	-8643,88	0,039	0,057	Model B vs. C	1	<0,0001	Neofunctionalization
		C : tree ratio	-8631,77	0,039	0,038	0,077			
D	6	A : One ratio	-7304,04	0,047		Model A vs. B	1	<0,05	
		B : two ratio	-7301,12	0,036	0,053	Model B vs. C	1	<0,001	Neofunctionalization
		C : tree ratio	-7295,66	0,036	0,040	0,068			
E	8	A : One ratio	-8608,62	0,044		Model A vs. B	1	<0,0001	
		B : two ratio	-8594,53	0,021	0,056	Model B vs. C	1	<0,0001	Neofunctionalization
		C : tree ratio	-8552,25	0,021	0,104	0,026			
	9	A : One ratio	-11100,46	0,029		Model A vs. B	1	<0,001	
		B : two ratio	-11094,96	0,020	0,032	Model B vs. C	1	NS	Relaxed purifying selection in both paralogues
		C : tree ratio	-11095,11	0,020	0,033	0,031			
	10	A : One ratio	-9970,81	0,051		Model A vs. B	1	<0,05	
		B : two ratio	-9968,20	0,040	0,056	Model B vs. C	1	<0,05	Neofunctionalization
		C : tree ratio	-9964,26	0,040	0,044	0,066			

Partie II. Duplication de génome et évolution lignée-spécifique : *sox11* chez le poisson-zèbre et le platy

Notre étude précédente démontre que la duplication de génome survenue à la base des poissons téléostéens il y a environ 300 Ma (Van de Peer *et al.* 2009) est à l'origine de l'expansion de la famille des gènes *sox* dans ce vaste groupe de vertébrés. Cette étude révèle également une évolution lignée-spécifique du contenu et de l'expression des gènes *sox* dans différentes espèces de poissons téléostéens. Par exemple, le gène *sox11* est préservé en deux copies dans les génomes de certains poissons téléostéens, dont le poisson-zèbre et le platy, et présent en singleton dans les génomes d'autres poissons téléostéens, tels que le médaka, le tétraodon, et le fugu. Les modèles classiques pour expliquer la préservation sur le long terme des gènes dupliqués sont la sous-fonctionnalisation, qui consiste en la distribution des différentes sous-fonctions du gène ancestral entre les deux paralogues, et/ou la néo-fonctionnalisation, où un paraglogue conserve la fonction ancestrale, tandis que le second acquiert une nouvelle fonction. Le premier objectif de cette seconde étude est de préciser les domaines d'expression des deux paralogues *sox11* par hybridation *in situ* dans le cerveau et les ovaires des poisson-zèbre et platy pour apporter une première indication sur le rôle biologique de ces gènes dans ces deux tissus où ils sont majoritairement exprimés. Dans un second temps, ces données sont destinées à fournir des indications sur les mécanismes évolutifs (sous-fonctionnalisation et/ou néo-fonctionnalisation) qui ont pu contribuer à la préservation des deux copies *sox11* dans chacun des deux poisson-zèbre et platy.

DIVERGENCE LIGNEE-SPECIFIQUE DES PARALOGUES *SOX11* APRES DUPLICATION DE GENOME

Résumé

Les poissons téléostéens, qui comptent autant d'espèces que l'ensemble des autres lignées de vertébrés réunies, constituent un groupe caractérisé par une très importante diversité morphologique, écologique et comportementale. La duplication de génome survenue chez l'ancêtre commun des téléostéens il y a environ 300 Ma, et l'enrichissement subséquent des répertoires de gènes, peuvent avoir facilité la diversification de ce vaste groupe d'animaux vertébrés. Cependant, les conséquences de la duplication de génome sur la biodiversité des téléostéens restent sujettes à débat dans la mesure où plus de 150 millions d'années séparent la duplication de génome de la radiation évolutive des téléostéens. Néanmoins, nous pouvons faire l'hypothèse que les gènes dupliqués, qui survivent sur de très longues périodes évolutives, peuvent évoluer indépendamment dans différentes lignées de poissons téléostéens et favoriser la diversification des espèces longtemps après la duplication de génome. Pour tester cette hypothèse, nous avons étudié l'évolution des domaines d'expression des deux gènes *sox11* chez le poisson-zèbre et le platy qui appartiennent respectivement aux deux grands groupes des otocéphales et des eutélostéens. Bien que les paralogues *sox11* semblent impliqués dans la neurogénèse et l'ovogénèse chez les deux espèces, nous avons observé une distribution des domaines d'expression entre les deux paralogues *sox11* différentes dans les deux lignées, ce qui suggère une évolution lignée-spécifique des paralogues après la divergence des otocéphales et des eutélostéens. Dans la mesure où la divergence de ces deux groupes est survenue longtemps après la duplication de génome (plusieurs dizaines de millions d'années), cette étude supporte l'hypothèse que les gènes dupliqués peuvent suivre de nouvelle trajectoire évolutive sur de longues périodes évolutives après duplication.

Introduction

La famille Sox (SRY-related HMG box) est une ancienne famille de facteurs de transcription présente chez tous les métazoaires et nommée après la découverte du gène du déterminisme du sexe *Sry* chez l'homme et la souris en 1990 (Gubbay *et al.* 1990, Sinclair *et al.* 1990). Les facteurs Sox possèdent un domaine HMG de 79 résidus qui permet la fixation à l'ADN. Chez les vertébrés, la famille Sox compte 19 membres (le groupe A, qui contient le vingtième membre *Sry*, est spécifique des mammifères). Les facteurs Sox sont des régulateurs transcriptionnels qui gouvernent la prolifération et la différenciation des cellules souches embryonnaires et adultes, et qui sont par conséquent impliqués dans le développement et l'homéostasie des vertébrés (Sarkar & Hochedlinger 2013). Le facteur de transcription Sox11 appartient au groupe C (avec Sox4 et Sox12) de la famille Sox. Chez les mammifères, les Sox du groupe C participent de manière redondante au développement embryonnaire de très nombreux organes, et notamment du cerveau (Dy *et al.* 2008, Bhattaram *et al.* 2010). Ils participent en effet au neurodéveloppement en contrôlant la différenciation des cellules précurseurs neuronales (Hoser *et al.* 2008, Bhattaram *et al.* 2010, Thein *et al.* 2010, Bergsland *et al.* 2011). L'expression des facteurs Sox11 au niveau du cerveau embryonnaire des téléostéens et des oiseaux suggère un rôle conservé des SoxC dans le développement du système nerveux central des vertébrés (De Martino *et al.* 2000, Kamachi *et al.* 2000, Uwanogho *et al.* 1995). Par ailleurs, chez l'adulte, les facteurs SoxC sont exprimés presque exclusivement dans les deux niches neurogéniques du cerveau des mammifères (le gyrus dentelé de l'hippocampe et le bulbe olfactif ; Zhao *et al.* 2008) où ils initient la différenciation des précurseurs neuronaux (Mu *et al.* 2012). Cependant, les fonctions des SoxC dans la neurogenèse adulte des autres animaux vertébrés ne sont pas connues. La famille des Sox est également impliquée dans la biologie de la reproduction des vertébrés, en particulier dans le développement des gonades (Schartl 2004, Santos *et al.* 2007). Par exemple, le gène du déterminisme du sexe *Sry*, mentionné plus haut, code la protéine *Sry* qui cible un amplificateur spécifique du gène *Sox9*, et conduit à la différenciation des testicules au cours du développement des mammifères (Sekido & Lovell-Badge 2008). Les SoxC, en particulier *Sox11*, sont exprimés chez certains poissons téléostéens dans les gonades femelles adultes, notamment chez le poisson-zèbre, la truite, et le mérrou, mais la fonction de ces gènes dans les ovaires est encore inconnue

(Santos *et al.* 2007, Kanda *et al.* 1998, Zhang *et al.* 2008, Zhang *et al.* 2010, Voldoire *et al.* en préparation).

Les deux copies du gène *sox11* (nommées *sox11a* et *sox11b*) identifiées dans différents génomes de poissons téléostéens sont la conséquence de la duplication de génome survenue chez l'ancêtre commun aux téléostéens (Hoegg *et al.* 2004, Voldoire *et al.* en préparation). En effet, les poissons téléostéens partagent un ancêtre commun polyploïde, et les génomes des poissons téléostéens contiennent encore de très nombreux gènes dupliqués (Amores *et al.* 1998, Jaillon *et al.* 2004, Kasahara *et al.* 2007). Les nouveaux gènes peuvent ensuite être recrutés pour de nouvelles fonctions, parfois longtemps après la duplication de génome, et promouvoir la diversification phénotypique et la spéciation (Conant & Wolfe 2008, Bridgham *et al.* 2008, Van de Peer *et al.* 2009). Nos travaux précédents montrent que la copie *sox11a* est présente dans tous les génomes des poissons téléostéens analysés et très conservée avec le gène orthologue *sox11* du lépisosté tacheté, du coelacanthé et des tétrapodes (figure 1 ; Voldoire *et al.* en préparation). A l'inverse, la copie *sox11b* présente une histoire évolutive plus complexe. En effet, le gène *sox11b* identifié chez les téléostéens non-percomorphes (paiche, poisson-zèbre, saumon, truite, morue) est également bien conservé avec l'orthologue *sox11* des espèces non-téléostéens et son paralogue *sox11a*. En revanche, le gène *sox11b* identifié chez certains percomorphes (tilapia, platy, guppy, épinoche et mériu) est beaucoup plus divergent : la reconstruction phylogénétique indique en effet une évolution moléculaire relativement rapide du gène *sox11b* chez l'ancêtre commun aux percomorphes. En outre, chez d'autres poissons percomorphes, le paralogue *sox11b* est absent du génome (médaka, tétraodon et fugu). Ces données *in silico* suggèrent la perte ou la préservation (par sous-fonctionnalisation et/ou néo-fonctionnalisation) lignée-spécifique de la copie *sox11b* qui a évolué rapidement à la base des percomorphes. L'étude des profils d'expression par qPCR des deux paralogues *sox11* chez le poisson-zèbre et le platy révèle que *sox11a* est majoritairement exprimé dans le cerveau des deux poissons, tandis que *sox11b* est principalement exprimé dans les ovaires (figure 2, Voldoire *et al.* en préparation). Il faut noter que chez le poisson-zèbre, seule la copie *sox11b* est exprimée dans les ovaires, tandis que les deux paralogues sont exprimés dans l'ovaire du platy. Ces données suggèrent un rôle biologique des paralogues *sox11* dans le cerveau et les ovaires des deux espèces. L'évolution moléculaire rapide du *sox11b* chez le platy peut être la signature d'une évolution des propriétés fonctionnelles de ce gène dans ces deux tissus.

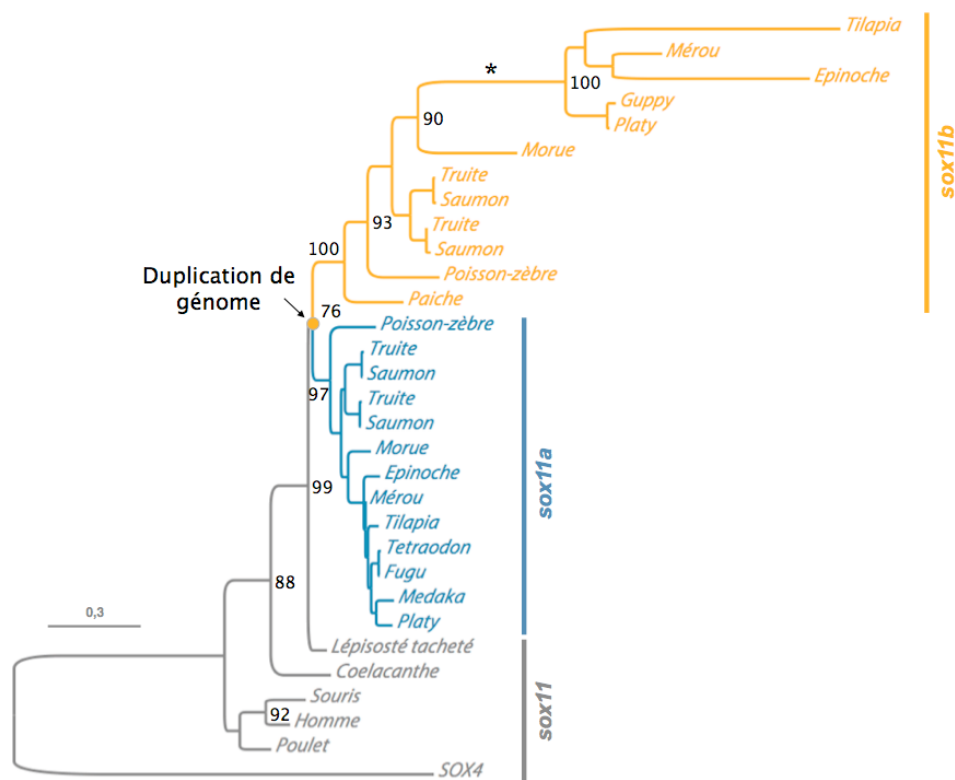


Figure 1 : Phylogénie du gène *sox11*. La reconstruction phylogénétique par la méthode du maximum de vraisemblance est basée sur la séquence protéique complète (modèle LG). L'arbre phylogénétique est enraciné avec la protéine SOX4 humaine qui constitue le groupe externe. La robustesse des branches de l'arbre est calculée avec le test statistique aLRT et indiquée pour les valeurs supérieures à 70. La duplication de génome survenue à la base des téléostéens est à l'origine des deux copies *sox11a* (branches bleues) et *sox11b* (branches oranges). Les orthologues non-téléostéens sont en singleton (branches grises). La longue branche à la base du clade qui regroupe les *sox11b* des poissons percomorphes est marquée d'une étoile.

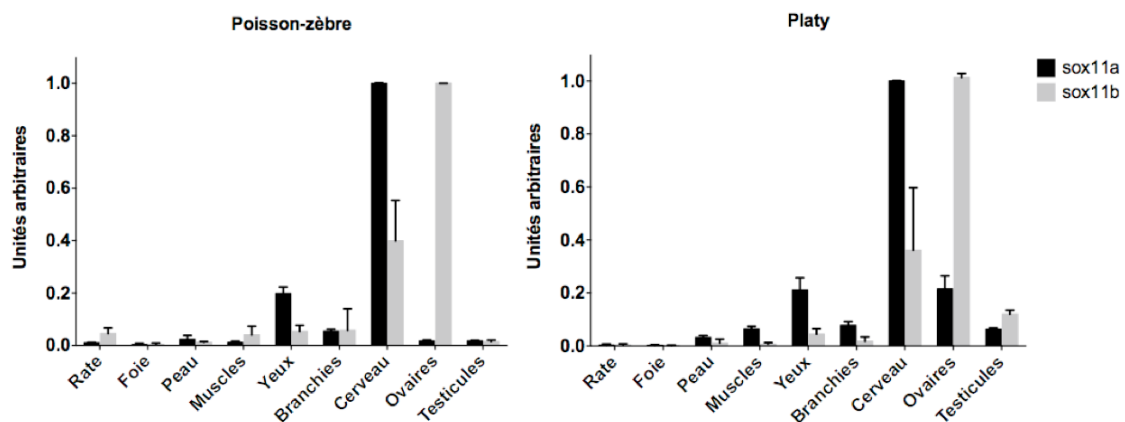


Figure 2 : Détection de l'expression par qPCR des deux paralogues *sox11* chez le poisson-zèbre et le platy adulte. La quantification relative des expressions de chaque paraglogue dans les divers tissus est calculée par la méthode des delta-delta-Ct.

Dans ce cadre, nous avons décidé de préciser le rôle biologique des deux *sox11* par hybridation *in situ* dans le cerveau et les ovaires des deux poissons téléostéens. En effet, la localisation précise des ARNm des deux paralogues dans le cerveau peut fournir une première indication sur un rôle potentiel conservé des *sox11* dans la neurogenèse adulte des deux poissons téléostéens. De même, nous avons décidé de localiser les ARNm des *sox11* dans les ovaires pour fournir de nouvelles indication sur leur fonction ovarienne (expression dans les cellules germinales et/ou somatiques). De plus, dans la mesure où les deux paralogues *sox11* sont exprimés dans l’ovaire du platy, l’hybridation *in situ* des ARNm correspondant peut permettre de déterminer si les deux gènes sont exprimés dans les mêmes cellules ou dans des cellules différentes, et en cela inférer une redondance ou une divergence fonctionnelle. Enfin, la comparaison des domaines d’expression obtenus dans le cerveau et les ovaires des poisson-zèbre et platy est utile pour tenter de retracer l’histoire évolutive des paralogues *sox11* chez ces deux espèces.

Matériels et méthodes

Synthèse des sondes pour l'hybridation in situ

Les deux paralogues *sox11a* et *sox11b* étant similaires au niveau de la séquence codante et divergentes au niveau des séquences 3'UTR et 5'UTR, les amorces ont été dessinées pour obtenir des sondes spécifiques qui s'hybrident sur environ 1000 pb à la fois sur la séquence codante (au niveau des régions les moins similaires) et sur la séquence 3'UTR. Les séquences des *sox11a* et *sox11b* choisies pour la synthèse des sondes ont été amplifiées par RT-PCR à partir des ARNm du cerveau et des ovaires pour chaque espèce, et clonées dans le vecteur pCR®II-TOPO® (Invitrogen). La séquence clonée dans le vecteur a été amplifiée avec les amorces universelles M13, puis purifiée avec le kit de purification PureLink™PCR (Invitrogen) et utilisée comme matrice pour la synthèse des sondes sens et anti-sens. Les sondes ont été synthétisées avec le kit DIG RNA labelling kit (SP6/T7, Roche) qui contient des bases uraciles étiquetées avec la digoxigénine (DIG) et les polymérases SP6 et T7. Les sondes ont été ensuite purifiées avec le kit RNesay (QIAGEN). La qualité a été vérifiée sur gel d'agarose et la quantité mesurée au Nanodrop (Shimadzu). Les sondes des *sox11a* et *sox11b* du poisson-zèbre font respectivement 1200 pb et 1000 pb, et celles du platy 1000 pb et 900 pb.

Préparation des coupes

Les poissons-zèbre (*Danio rerio*, lignée Jp163A) et platy (*Xiphophorus maculatus*, population Rio Jamapa WLC1274) adultes femelles âgées de 8 mois ont été fixés avec du paraformaldéhyde 4% dilué dans du tampon phosphate salin (PBS, pH 7,4) à 4°C pendant 20 heures. Puis, les poissons ont été lavés dans du PBS pendant 2 heures, coupés en deux tronçons (tête et abdomen) et cryo-protégés par incubation dans du PBS-saccharose (30%) à 4°C sur la nuit. Les échantillons ont été rapidement congelés dans du milieu Tissue-Tek® O.C.T.™ et conservés à -80°C. Les échantillons ont été sectionnés avec une lame en carbure de tungstène dans un cryostat (Leica CM3050S) en coupes de 16 µm montées ensuite sur lame SuperFrost (Fisher Scientific, Atlanta, GA).

Hybridation in situ

L'hybridation *in situ* a été réalisée en suivant le protocole décrit dans (Schaeren-Wiemers & Gerfin-Moser 1993) avec quelques modifications. Les coupes ont été séchées à température ambiante pendant 20 minutes, fixées au paraformaldéhyde 4% dilué dans

du PBS, lavées trois fois avec du PBS-Tween, et incubées dans du méthanol (3 % d'eau oxygénée) pendant 30 min. Après lavage, les coupes ont été acétylées pendant 10 min puis à nouveau lavées. La pré-hybridation été réalisée avec le tampon d'hybridation [50% formamide, citrate de sodium salin 5X (SSC), solution Denhardt's 5X (Euromedex), 200 µg/ml de tRNA de levure (Roche) et 500 µg/ml d'ADN de sperme de saumon (Sigma)] pendant 2 heures à 68°C en chambre humidifiée avec du SSC 5X. Les coupes ont été ensuite hybridées pendant la nuit à 68°C avec du tampon d'hybridation additionné de la sonde (sens ou anti-sens). Les coupes ont été lavées et incubées dans du SSC 5X pendant 30 minutes à 68°C, puis dans du SSC 0,2X trois fois 30 minutes à 68°C. Les coupes ont été ensuite rincées avec du MAB-T (tampon tris-maléate contenant 0,01% de Tween20), bloquées avec le tampon de blocage [1% de l'agent de blocage (Roche) et 2% de sérum de mouton (Sigma) dans du MAB-T] et incubées une nuit à 4°C avec l'anticorps anti-phosphatase alcaline anti-DIG (Roche) diluée au 1/5000 dans du tampon de détection [0,5% de l'agent de blocage et 1% de sérum de mouton dans du MAB-T]. Les coupes ont été lavées avec du PBS, puis le signal révélé avec du NBT-BCIP (Fermentas). Finalement, les coupes ont été montées dans du milieu Aquamount (Polysciences, Inc).

Résultats et discussion

Expression des sox11 dans le cerveau et les ovaires de deux poissons téléostéens adultes

Nous avons localisé les ARNm des deux paralogues *sox11a* et *sox11b* dans le cerveau des téléostéens adultes par hybridation *in situ* sur des coupes transversales réalisées le long de l'axe antéro-postérieur du cerveau (figure 3a, b, d, e, f). Les coupes histologiques sur la figure 3 présentent la majorité des structures du cerveau où les gènes *sox11a* et *sox11b* sont exprimés. Ainsi, chez le poisson-zèbre, les deux paralogues sont exprimés dans le mésencéphale (figure 3, a1, b2) dans une mince couche cellulaire logée entre le toit optique et le cervelet (nommée *posterior mesencephalic lamina*, PML ; Grandel *et al.* 2006) et dans la zone périventriculaire du toit optique (*periventricular gray zone of optic tectum*, PGZ ; Zupanc *et al.* 2005). Dans le diencéphale (figure 3, a2, b1), les ARNm des deux paralogues *sox11a* et *sox11b* sont respectivement détectés dans les subdivisions dorsale et caudale de la zone périventriculaire de l'hypothalamus (*periventricular dorsal hypothalamus*, Hd et *periventricular caudal hypothalamus*, Hc ; Grandel *et al.* 2006). Enfin, chez le poisson-zèbre, le gène *sox11b* est exprimé dans le diencéphale (TLa et DIL, figure 3, b1), dans le rhombencéphale (figure 3, b2), au niveau de la couche granulaire du cervelet (*granular layer of cerebellar corpus*, CCe_{gra} ; Zupanc *et al.* 2005), et dans le télencéphale (figure 3c), au niveau du pallium (Mueller 2012). Les différentes régions où sont localisés les ARNm des deux paralogues *sox11a* et *sox11b* sont décrites chez le poisson-zèbre pour être des zones neurogéniques (Zupanc *et al.* 2005, Grandel *et al.* 2006, Mueller 2012, Kizil *et al.* 2012). Il faut noter ici que le cerveau des téléostéens adultes est caractérisé par la présence de nombreuses zones neurogéniques où les cellules souches neuronales prolifèrent et se différencient en cellules neurales (Zupanc *et al.* 2005) et contraste en cela avec celui des mammifères où la neurogenèse est restreinte chez l'adulte à deux régions du cerveau : hippocampe et bulbe olfactif (Zhao *et al.* 2008). En outre, comme mentionné plus haut, le gène *Sox11* est exprimé dans le cerveau de la souris adulte presque exclusivement au niveau du gyrus dentelé de l'hippocampe et du bulbe olfactif, où il initie la différenciation des cellules souches neuronales (Mu *et al.* 2012). L'ensemble de ces données suggère que la fonction du facteur de transcription Sox11 dans la neurogenèse adulte est conservée entre le souris et le poisson-zèbre. Chez le platy, les deux paralogues sont exprimés dans le mésencéphale (figure 3, d1, f1, f2) au niveau d'une mince couche logée entre le toit optique et le cervelet qui semble homologue à celle décrite plus haut chez le poisson-

zèbre (*posterior mesencephalic lamina*, PML). Il faut noter que, chez le platy, *sox11b* n'est exprimé que dans cette zone restreinte du cerveau où son domaine d'expression est chevauchant avec celui de *sox11a*. En revanche, *sox11a* est exprimé également dans le mésencéphale (figure 3, d1) au niveau de la zone périventriculaire du toit optique (PGZ), dans le diencéphale (figure 3, d2), en particulier dans des régions qui semblent homologues à l'hypothalamus du poisson-zèbre, et également dans le rhombencéphale (figure 3, d1) au niveau du cervelet. Enfin, chez le platy, c'est *sox11a* qui est exprimé dans le télencéphale au niveau d'une zone homologue au pallium du poisson-zèbre (figure 3e, MP, SP, VP). Contrairement au poisson-zèbre, l'anatomie du cerveau du platy, ainsi que les différentes zones neurogéniques, ne sont pas décrites dans la littérature. Par conséquent, il est difficile de conclure à un rôle potentiel des paralogues dans la neurogenèse adulte chez le platy. Néanmoins, l'expression des paralogues chez le platy, en particulier *sox11a*, est localisée principalement dans certaines zones du cerveau qui semblent homologues aux zones neurogéniques du poisson-zèbre. Finalement, l'étude conjointe de l'expression des paralogues *sox11* et de marqueurs des cellules précurseurs neuronales, par exemple les gènes codant les protéines nucléaires neuronales Hu-C et Hu-D (Grandel *et al.* 2006), est nécessaire pour d'une part confirmer ou infirmer un rôle potentiel des *sox11* dans la neurogenèse chez le poisson-zèbre, et d'autre part déterminer si les paralogues *sox11* sont exprimés dans les zones neurogéniques chez le platy.

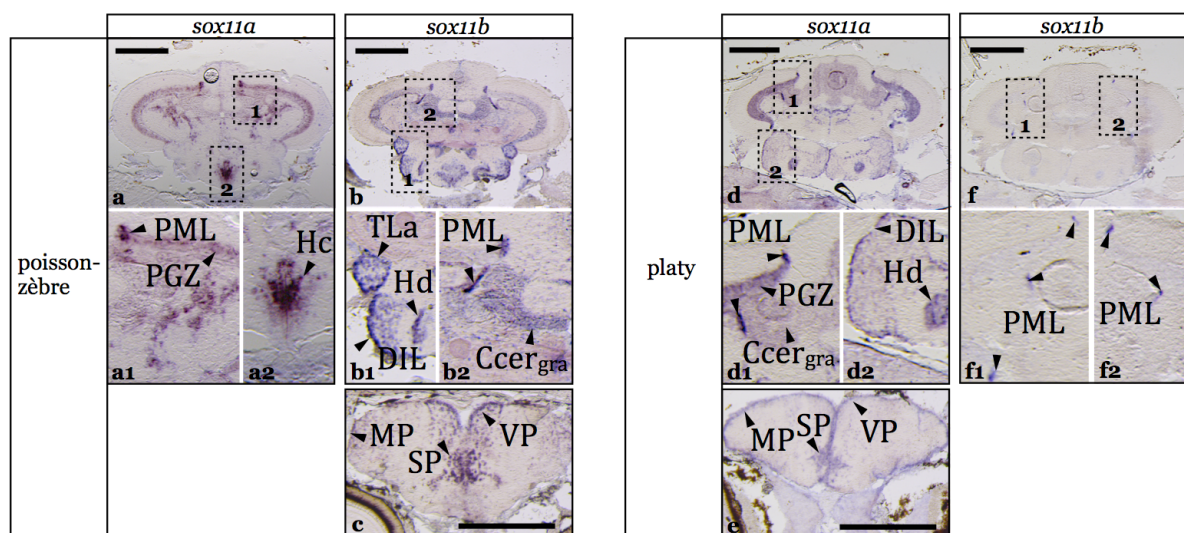


Figure 3 : Détection par hybridation *in situ* de l'expression des paralogues *sox11a* et *sox11b* dans le cerveau du poisson-zèbre et du platy. PML : *posterior mesencephalic lamina*, PGZ : *periventricular gray zone of optic tectum*, Hc : *periventricular caudal hypothalamus*, Hd : *periventricular dorsal hypothalamus*, TLa : *lateral torus*, Ccer_{gra} : *granular layer of cerebellar corpus*, DIL : *nucleus diffusus lobi inferioris hypothalami*, MP : *medial pallium*, VP : *ventral pallium*, SP : *sub-pallium*. Barre d'échelle : 500 μ m.

Nous avons également étudié l'expression des deux paralogues *sox11* dans les ovaires des poisson-zèbre et platy en réalisant des hybridations *in situ* sur des coupes transversales d'ovaires effectuées au niveau de l'abdomen (figure 4). Chez le poisson-zèbre, le gène *vasa* (figure 4a, d), qui est un marqueur des cellules germinales, est exprimé dans les ovogonies, autour de la vésicule germinative (noyau de l'ovocyte) dans certains oocytes précoces de stade I, ainsi qu'au niveau du cortex végétatif dans des ovocytes un peu plus tardifs de stade I et de stade II. Enfin, les ARNm du gène *vasa* sont localisés au niveau du cortex végétatif dans les oocytes de stade III (au sein desquels les granules vitellins commencent à s'accumuler). Ces observations sont en accord avec les travaux effectués préalablement par Kosaka *et al.* (2007) qui précisent que les ARNm du gène *vasa* sont localisés au niveau du nuage mitochondrial, une région cytoplasmique riche en mitochondries, et précurseur du plasme germinatif (Kosaka *et al.* 2007). Chez le poisson-zèbre, comme observé par qPCR (figure 2), *sox11a* n'est pas exprimé dans les ovaires (figure 4b, e). En revanche, *sox11b* est exprimé à plusieurs stades de maturation des ovocytes, en particulier les stades I, II et III (figure 4c, f). Contrairement à *vasa*, les ARNm de *sox11b* sont localisés de manière homogène autour de la vésicule germinative dans les ovocytes de stade I et II puis dans l'ensemble du cytoplasme des ovocytes de stade III. Ces résultats suggèrent que *sox11b*, qui est fortement exprimé dans les ovogonies, mais également dans les premiers stades de maturation des oocytes, peut participer au renouvellement des ovogonies (qui persistent et continuent de se diviser tout au long de la vie chez le poisson-zèbre) et également être impliqué dans l'ovogenèse chez le poisson-zèbre. Chez le platy, le gène *vasa* semble également être exprimé, comme précédemment, dans les ovocytes précoces, de manière plus homogène autour de la vésicule germinale, et ainsi marquer les cellules germinales (figure 4g) (Potter & Kramer 2000). Les deux *sox11a* et *sox11b* (figure 4h, i) chez le platy sont exprimés dans les mêmes ovocytes que *vasa*, ce qui suggère que les deux paralogues jouent un rôle dans les cellules primordiales. L'expression des *sox11* dans les ovaires des deux platy et poisson-zèbre est en accord avec des études antérieures réalisées chez le poisson-zèbre, la truite et le mérrou, qui montrent également une expression ovarienne d'un ou des deux paralogues, et en cela un rôle potentiel dans l'ovogenèse (Santos *et al.* 2007, Kanda *et al.* 1998, Zhang *et al.* 2008, Zhang *et al.* 2010).

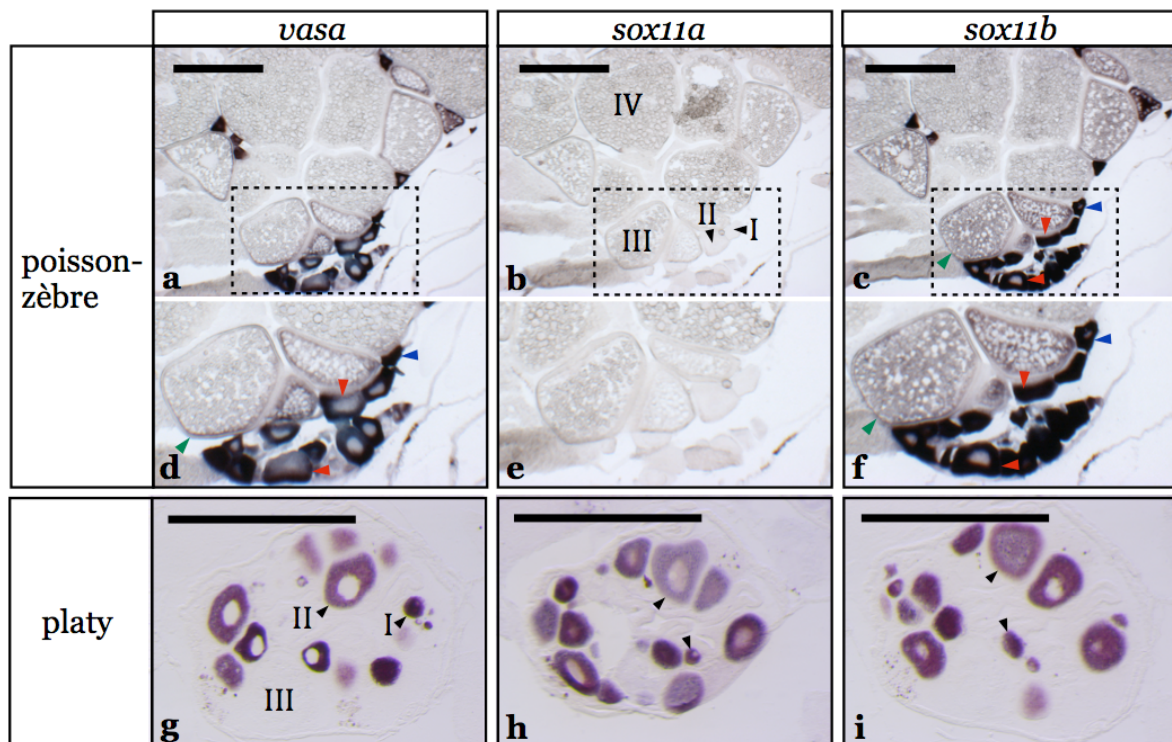


Figure 4 : Caractérisation par hybridation *in situ* de l'expression des paralogues *sox11a* et *sox11b* dans les ovaires du poisson-zèbre et du platy. Les ovocytes aux stades I, II et III sont indiqués sur la figure. Dans les ovaires du poisson-zèbre, l'expression dans les ovocytes I, II et III des *vasa* et *sox11b* est indiquée respectivement par une pointe bleu, rouge et verte. Barre d'échelle : 500 μ m.

*Histoire évolutive des paralogues *sox11* chez les téléostéens*

Dans cette étude, les premières indications sur les rôles fonctionnels potentiels des deux paralogues *sox11* obtenus chez le poisson-zèbre et le platy peuvent être mis à profit pour inférer l'histoire évolutive de ces deux gènes chez ces deux espèces (figure 5). Ainsi, ces résultats montrent que les deux paralogues sont exprimés dans le cerveau où ils présentent des profils d'expression différents chez le poisson-zèbre et le platy. Les expressions des *sox11a* et *sox11b* dans le cerveau du poisson-zèbre sont superposées dans certaines régions, par exemple au niveau du toit optique, et sont en revanche distinctes dans certaines autres régions au niveau de l'hypothalamus, du télencéphale et du mésencéphale. En outre, ces données montrent que les ARNm des deux paralogues sont localisés ou co-localisés dans les zones neurogéniques du cerveau du poisson-zèbre (Zupanc *et al.* 2005, Grandel *et al.* 2006, Mueller 2012). Ce constat suggère que les deux

gènes sont impliqués dans la neurogenèse adulte du poisson-zèbre, potentiellement de manière redondante au niveau de certaines zones, et distincte dans d'autres zones. Dans la mesure où le gène orthologue *Sox11* est également exprimé dans les zones neurogéniques du cerveau des mammifères (où il initie la différenciation des cellules précurseurs neurales ; Mu *et al.* 2012), nous pouvons supposer que le gène ancestral *Sox11*, présent chez l'ancêtre commun des téléostéens et des mammifères, avait un rôle dans la neurogenèse adulte (figure 5). Selon cette hypothèse, les deux gènes *sox11a* et *sox11b* chez le poisson-zèbre pourraient illustrer un cas de sous-fonctionnalisation, avec le partage des différents domaines d'expression ancestraux entre les deux paralogues. Il faut rappeler ici que les données *in silico* indiquent une forte conservation de la séquence primaire des deux paralogues *sox11* du poisson-zèbre entre eux (et avec les orthologues non-téléostéens), ce qui suggèrent des propriétés fonctionnelles conservées. En outre, notre précédente étude a mis en évidence le partage entre les deux paralogues *sox11* du poisson-zèbre de certains éléments très conservés entre vertébrés (HCNE, *highly conserved non-coding elements*) au niveau des séquences non-codantes (qui sont de potentiels éléments *cis*-régulateurs) (Voldoire *et al.* en préparation). Ainsi, l'ensemble de ces données supporte l'hypothèse d'une sous-fonctionnalisation des deux copies dans le cerveau du poisson-zèbre. Chez le platy, il en est tout autrement, et *sox11a* présente un spectre d'expression beaucoup plus large que *sox11b* (figure 5). De plus, les quelques zones très restreintes du toit optique où est exprimé *sox11b* sont superposées avec *sox11a*. Comme mentionné précédemment, le gène *sox11a* est quant à lui exprimé dans de nombreuses régions du cerveau du platy qui semblent homologues aux zones neurogéniques du poisson-zèbre. Il est intéressant de souligner ici que *sox11a* chez le platy est en effet exprimé dans les mêmes régions que son orthologue *sox11a* chez le poisson-zèbre (toit optique, hypothalamus notamment), mais également dans les régions qui expriment uniquement le gène *sox11b* chez le poisson-zèbre (au niveau du lobe inférieur de l'hypothalamus et du télencéphale). Selon l'hypothèse précédemment exposée, le paraglogue *sox11a* du platy aurait conservé la majorité des domaines d'expression ancestraux au niveau des zones neurogéniques dans le cerveau, contrairement au paraglogue *sox11b* qui aurait perdu la plupart des domaines d'expression ancestraux. En accord avec ces données d'expression, la grande majorité des HNCE conservés au voisinage du *sox11a* chez le platy sont perdus au voisinage du *sox11b* (Voldoire *et al.* en préparation). Enfin, il faut noter ici que la divergence entre les séquences primaires des deux paralogues *sox11* du platy est beaucoup plus importante que chez le poisson-zèbre, et la reconstruction phylogénétique indique en effet une

évolution moléculaire rapide du *sox11b* chez l'ancêtre commun aux percomorphes (figure 1). L'ensemble de ces données suggèrent une évolution des propriétés fonctionnelles du paralogue *sox11b* dans le cerveau du platy, éventuellement associée à une spécialisation de ce paralogue dans certaines fonctions cérébrales.

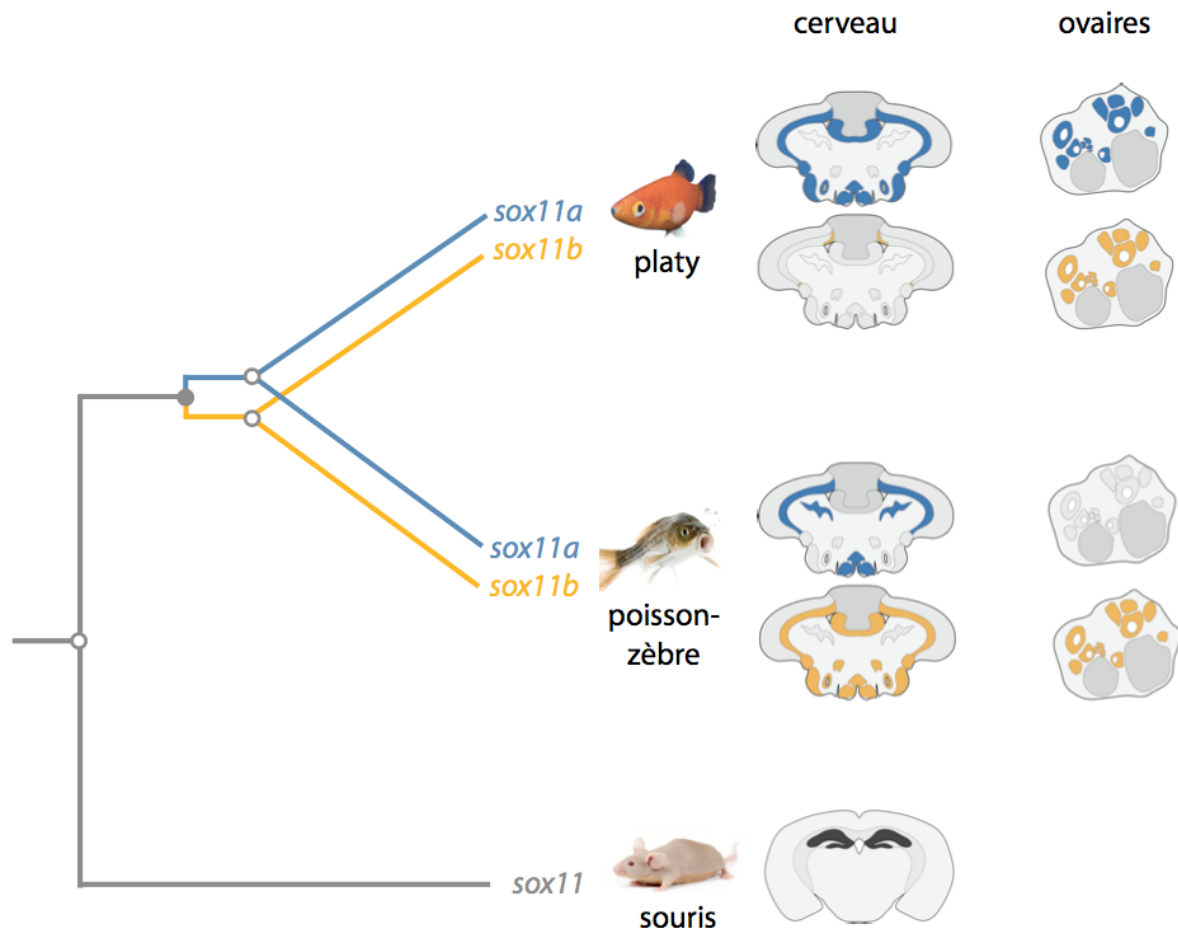


Figure 5 : Histoire évolutive du gène *sox11* chez trois vertébrés. Les ronds blancs indiquent les événements de spéciation et le rond gris indique la duplication du gène *sox11* chez les deux poissons téléostéens. Le gène *sox11* murin est exprimé dans l'hippocampe (gris foncé) et dans le bulbe olfactif (non représenté sur le schéma). Le gène *sox11a* est exprimé dans le cerveau des deux poisson-zèbre et platy et dans l'ovaire du platy (bleu). Le gène *sox11b* est exprimé dans le cerveau et les ovaires des deux poissons téléostéens. Cependant, le spectre d'expression du *sox11b* dans le cerveau du platy est très restreint. *Sox11* dans les ovaires de la souris n'est pas exprimé.

Par ailleurs, cette étude montre également une évolution des profils d'expression des deux gènes *sox11a* et *sox11b* dans les ovaires des poisson-zèbre et platy. Chez le poisson-zèbre, seul le paralogue *sox11b* est exprimé au niveau des cellules germinales des ovaires, tandis que les deux paralogues *sox11a* et *sox11b* sont exprimés dans les cellules primordiales chez le platy. Il faut noter ici que le gène *sox11b* chez le platy, qui est

exprimé dans des régions très restreintes du cerveau, présente en revanche une expression dans différents types d'ovocytes, ce qui suggère une spécialisation de ce paralogue dans les fonctions ovariennes. L'expression des paralogues *sox11* dans les ovaires est également observée chez d'autres espèces de poissons téléostéens, notamment chez la truite et le mérou, ce qui suggère une expression ovarienne du *sox11* ancestral chez l'ancêtre commun aux téléostéens (Kanda *et al.* 1998, Zhang *et al.* 2008, Zhang *et al.* 2010). Selon cette hypothèse, la fonction ovarienne est réalisée chez le poisson-zèbre par le paralogue *sox11b*. Chez le platy, la co-expression des deux paralogues dans les mêmes cellules de l'ovaire peut être associée à une redondance fonctionnelle. Néanmoins, la divergence de la séquence protéique du *sox11b* observée *in silico* peut être associée à de nouvelles interactions, et selon cette hypothèse, les deux paralogues, bien qu'exprimés tous les deux dans les mêmes ovocytes, peuvent avoir deux fonctions différentes.

Conclusion

Nous avons montré dans une précédente étude que les génomes de nombreux poissons téléostéens contiennent deux copies du gène *sox11*, issues de la duplication de génome complet survenue à la base des téléostéens (Voldoire *et al.* en préparation, Jaillon *et al.* 2004). La présente étude fournit d'une part de nouvelles indications sur les fonctions potentielles des gènes *sox11* chez les poissons téléostéens adultes, en particulier dans les tissus où ils sont majoritairement exprimés : le cerveau et les ovaires. Nous avons montré en effet par hybridation *in situ* que les ARNm des deux paralogues *sox11* chez les deux poisson-zèbre et platy sont localisés dans le cerveau au niveau des zones neurogéniques, et dans les ovaires au niveau des cellules primordiales. Ces données suggèrent un rôle des *sox11* dans la neurogénèse adulte et dans l'ovogénèse des poissons téléostéens. Ces conclusions sont en accord avec les fonctions attribuées aux facteurs de transcription Sox dans la prolifération et la différenciation des cellules souches adultes chez les vertébrés (Sarkar & Hochedlinger 2013). Néanmoins, l'étude de différents marqueurs de prolifération et de différenciation cellulaire, en particulier des cellules précurseurs neuronales et des cellules primordiales, peut permettre à l'avenir de comprendre plus amplement le rôle des *sox11* dans ces processus biologiques. D'autre part, cette étude fournit des indications sur l'évolution des paralogues *sox11* chez les poissons téléostéens, en particulier chez le poisson-zèbre et le platy. Ainsi, l'ensemble des résultats indique clairement une évolution espèce-spécifique des deux paralogues. En particulier, l'évolution rapide du paraglogue *sox11b* chez l'ancêtre commun aux percomorphes est certainement responsable de la divergence fonctionnelle de ce paraglogue dans le cerveau du platy (tandis que les profils d'expression des deux *sox11a* et *sox11b* du poisson-zèbre supposent plutôt un cas de sous-fonctionnalisation). En outre, on ne peut pas exclure un cas de néo-fonctionnalisation du *sox11b* divergent dans l'ovaire du platy (le *sox11a* réaliserait alors la fonction ovarienne ancestrale). Ainsi, l'évolution rapide du gène *sox11b* chez les percomorphes peut être la conséquence d'un relâchement de la sélection négative, avec accumulation de mutations dégénératives, qui aurait abouti chez certaines espèces à une perte par pseudogénisation de ce gène (par exemple chez le médaka, le fugu et le tétraodon), et chez certaines autres espèces à une sous-fonctionnalisation et/ou néo-fonctionnalisation (par exemple chez le platy et le méné). Finalement, cette étude supporte l'idée que les événements de non-fonctionnalisation et/ou néo-fonctionnalisation peuvent survenir longtemps après

l'événement de duplication de gènes, dans le cas présent très certainement après la divergence des otocéphales et des eutélostéens, chez l'ancêtre des percomorphes, quelque 200 millions d'années après la duplication de génome spécifique aux téléostéens. Ces mécanismes génétiques, appliqués à un très grand nombre de gènes dupliqués, peuvent en cela avoir facilité la diversification des phénotypes et la radiation des téléostéens sur de longues périodes évolutives après la duplication de génome survenue à la base des téléostéens.

Perspectives

*Caractérisation fine des populations cellulaires exprimant les gènes *sox11**

Les conclusions préliminaires de cette étude ouvrent de nombreuses perspectives pour comprendre le rôle biologique et l'évolution des gènes *sox11a* et *sox11b* chez le poisson-zèbre et le platy. D'une part, nous devons confirmer le rôle potentiel des gènes *sox11* dans la neurogenèse adulte des deux poissons téléostéens. Pour cela, nous devons déterminer si les cellules qui expriment *sox11a* et *sox11b* (au niveau des zones neurogéniques) sont bien des cellules souches neuronales adultes en prolifération et/ou des cellules précurseurs neuronales en cours de différenciation. Le Bromo-desoxyuridine (BrdU) qui marque les cellules en mitose, et l'utilisation des marqueurs des cellules souches neuronales tels que *pcna* (*proliferating cell nuclear antigen*), *sox2* et *msi1* (*musashi homolog 1*) permettront d'identifier les cellules souches neuronales en prolifération dans le cerveau des deux poissons téléostéens (Ito *et al.* 2010). Les marqueurs des cellules neuronales en cours de différenciation *Hu-C* et *Hu-D*, comme mentionné plus haut, permettront d'identifier quant à eux les cellules précurseurs neuronales en cours de différenciation (Grandel *et al.* 2006). Ces données permettront également d'identifier une potentielle spécialisation des paralogues *sox11* dans le maintien des cellules souches ou la différenciation des cellules précurseurs neuronales dans le cerveau des deux téléostéens. Par exemple, nous pourrions tester l'hypothèse d'une spécialisation du gène *sox11b* du platy dans le maintien des cellules souches neuronales. En effet, ce gène est exprimé dans une région restreinte du cerveau du platy qui semble homologue à une zone de prolifération chez le poisson-zèbre (figure 3, Ito *et al.* 2010).

Caractérisation des capacités de trans-activation des différentes protéines Sox11

La divergence fonctionnelle du gène *sox11b* du platy est suggérée dans cette étude par une divergence à la fois des séquences codante et non-codante (*in silico*) et du profil d'expressions de ce gène, en particulier dans le cerveau. Cependant, ces données indiquent seulement une partie de l'histoire évolutive de ce gène chez le platy. En effet, l'évolution fonctionnelle des facteurs de transcription est médiée en partie par une évolution des séquences *cis*-régulatrices (suggérée dans cette étude), et en partie par une évolution des propriétés biochimiques de la protéine (spécificité de fixation à l'ADN, nouveaux partenaires d'interaction, évolution de la régulation post-traductionnelle) (Escriva *et al.* 2006). L'étude comparative des propriétés biochimiques des protéines codées par les gènes *sox11a* et *sox11b* des deux poissons téléostéens est envisagée pour tester l'hypothèse d'une évolution des propriétés biochimiques du *sox11b* chez le platy. Le protocole pour réaliser cette expérience est inspiré d'un travail effectué par Lefebvre et son équipe sur la caractérisation des propriétés moléculaires des SoxC (Sox4, Sox11 et Sox12) chez la souris (Dy *et al.* 2008). Les auteurs utilisent le système rapporteur luciférase pour mesurer la capacité de trans-activation des protéines SoxC. Leur plasmide rapporteur contient en amont du gène de la luciférase le promoteur proximal du gène *Tubb3* de la souris. Les travaux de Bergsland *et al.* 2006 montrent en effet que le gène de la tubuline *Tubb3* chez la souris est une cible directe des facteurs Sox4 et Sox11 au cours du neurodéveloppement de la souris. Ainsi, Lefebvre et son équipe (2008) ont réalisé une co-transfection transitoire dans des cellules Cos1 avec le plasmide d'expression contenant *Sox4*, *Sox11* ou *Sox12* et le plasmide rapporteur décrit précédemment. Ils montrent des différences dans les capacités de transactivation entre les différents facteurs SoxC (Dy *et al.* 2008). Ainsi, nous pourrions reprendre ce protocole pour tester les capacités de trans-activation des protéines Sox11a et Sox11b du poisson-zèbre et du platy. Ces données donneront des indications sur les différences dans les capacités de trans-activation entre protéines paralogues et orthologues Sox11, et permettront potentiellement de relier la divergence de la séquence primaire du Sox11b du platy à une évolution des capacités de trans-activation.

Références

- Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y., et al. (1998). Zebrafish hox clusters and vertebrate genome evolution. *SCIENCE* 282, 1711–1714.
- Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., and Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.* 25, 2453–2464.
- Bhattaram, P., Penzo-Mendez, A., Sock, E., Colmenares, C., Kaneko, K., Vassilev, A., DePamphilis, M., Wegner, M., and Lefebvre, V. (2010). Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nat. Commun.* 1.
- Bridgham, J.T., Brown, J.E., Rodríguez-Marí, A., Catchen, J.M., and Thornton, J.W. (2008). Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors. *PLoS Genet* 4, e1000191.
- Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernández, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Semin. Cell Dev. Biol.* 24, 83–94.
- Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* 9, 938–950.
- Dy, P., Penzo-Méndez, A., Wang, H., Pedraza, C.E., Macklin, W.B., and Lefebvre, V. (2008). The three SoxC proteins—Sox4, Sox11 and Sox12—exhibit overlapping expression patterns and molecular properties. *Nucleic Acids Res.* 36, 3101–3117.
- Grandel, H., Kaslin, J., Ganz, J., Wenzel, I., and Brand, M. (2006). Neural stem cells and neurogenesis in the adult zebrafish brain: origin, proliferation dynamics, migration and cell fate. *Dev. Biol.* 295, 263–277.
- Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* 346, 245–250.
- Hiraoka, Y., Komatsu, N., Sakai, Y., Ogawa, M., Shiozawa, M., and Aiso, S. (1997). XLS13A and XLS13B: SRY-related genes of *Xenopus laevis*. *GENE* 197, 65–71.
- Hoser, M., Potzner, M., Koch, J., Bosl, M., Wegner, M., and Sock, E. (2008). Sox12 deletion in the mouse reveals nonreciprocal redundancy with the related Sox4 and Sox11 transcription factors. *Mol. Cell. Biol.* 28, 4675–4687.
- Ito, Y., Tanaka, H., Okamoto, H., Ohshima, T. (2010). Characterization of neural stem cells and their progeny in the adult zebrafish optic tectum. *Dev. Biol.* 342, 26–38.
- Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *NATURE* 431, 946–957.
- Kamachi, Y., Uchikawa, M., and Kondoh, H. (2000). Pairing SOX off: with partners in the regulation of embryonic development. *Trends Genet.* TIG 16, 182–187.
- Kanda, H., Kojima, M., Miyamoto, N., Ito, M., Takamatsu, N., Yamashita, S., and Shiba, T. (1998). Rainbow trout Sox24, a novel member of the Sox family, is a transcriptional regulator during oogenesis. *GENE* 211, 251–257.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719.
- Kizil, C., Kaslin, J., Kroehne, V., and Brand, M. (2012). Adult neurogenesis and brain regeneration in zebrafish. *Dev. Neurobiol.* 27, 429–461.

- Kosaka, K., Kawakami, K., Sakamoto, H., and Inoue, K. (2007). Spatiotemporal localization of germ plasm RNAs during zebrafish oogenesis. *Mech. Dev.* *124*, 279–289.
- Lynch, M., and Force, A.G. (2000). The origin of interspecific genomic incompatibility via gene duplication. *Am. Nat.* *156*, 590–605.
- De Martino, S., Yan, Y., Jowett, T., Postlethwait, J., Varga, Z., Ashworth, A., and Austin, C. (2000). Expression of *sox11* gene duplicates in zebrafish suggests the reciprocal loss of ancestral gene expression patterns in development. *Dev. Dyn.* *217*, 279–292.
- Mu, L., Berti, L., Masserdotti, G., Covic, M., Michaelidis, T.M., Doberauer, K., Merz, K., Rehfeld, F., Haslinger, A., Wegner, M., et al. (2012). SoxC Transcription Factors Are Required for Neuronal Differentiation in Adult Hippocampal Neurogenesis. *J. Neurosci.* *32*, 3067–3080.
- Mueller, T. (2012). What is the Thalamus in Zebrafish? *Front. Neurosci.* *6*, 64.
- Van de Peer, Y., Maere, S., and Meyer, A. (2009). OPINION The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* *10*, 725–732.
- Potter, H., and Kramer, C.R. (2000). Ultrastructural observations on sperm storage in the ovary of the platyfish, *Xiphophorus maculatus* (Teleostei: poeciliidae): the role of the duct epithelium. *J. Morphol.* *245*, 110–129.
- Prince, V. E. and Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. (2002). *Nat. Rev. Genet.* *3*, 827–837.
- Santos, E.M., Workman, V.L., Paull, G.C., Filby, A.L., Look, K.J.W.V., Kille, P., and Tyler, C.R. (2007). Molecular basis of sex and reproductive status in breeding zebrafish. *Physiol. Genomics* *30*, 111–122.
- Sarkar, A., and Hochedlinger, K. (2013). The *sox* family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell* *12*, 15–30.
- Schaeren-Wiemers, N., Gerfin-Moser, A. (1993) A single protocol to detect transcripts of various types and expression levels in neural tissue and cultured cells: *in situ* hybridization using digoxigenin-labelled cRNA probes. *Histochemistry*. *100*, 431–40.
- Schartl, M. (2004). Sex chromosome evolution in non-mammalian vertebrates. *Curr. Opin. Genet. Dev.* *14*, 634–641.
- Sekido, R. & Lovell-Badge, R. (2008). Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer. *Nature* *453*, 930–934.
- Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischau, A.M., Lovell-Badge, R., and Goodfellow, P.N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* *346*, 240–244.
- Thein, D., Thalhammer, J., Hartwig, A., Crenshaw, E., Lefebvre, V., Wegner, M., and Sock, E. (2010). The closely related transcription factors *Sox4* and *Sox11* function as survival factors during spinal cord development. *J. Neurochem.* *115*, 131–141.
- Uwanogho, D., Rex, M., Cartwright, E., Pearl, G., Healy, C., Scotting, P., and Sharpe, P. (1995). Embryonic expression of the chicken *Sox2*, *Sox3*, and *Sox11* genes suggests an interactive role in neuronal development. *Mech. Dev.* *49*, 23–36.
- Volff, J. (2005). Genome evolution and biodiversity in teleost fish. *HEREDITY* *94*, 280–294.
- Zhang, L., Lin, D., Zhang, Y., Ma, G., and Zhang, W. (2008). A homologue of *Sox11* predominantly expressed in the ovary of the orange-spotted grouper *Epinephelus coioides*. *Comp. Biochem. Physiol. B-Biochem. Mol.* *149*, 345–353.

Zhang, L., Zhu, T., Lin, D., Zhang, Y., and Zhang, W. (2010). A second form of Sox11 homologue identified in the orange-spotted grouper *Epinephelus coioides*: Analysis of sequence and mRNA expression patterns. *Comp. Biochem. Physiol. B-Biochem. Mol.* *157*, 415–422.

Zhao, C., Deng, W., and Gage, F.H. (2008). Mechanisms and Functional Implications of Adult Neurogenesis. *Cell* *132*, 645–660.

Zupanc, G.K.H., Hinsch, K., and Gage, F.H. (2005). Proliferation, migration, neuronal differentiation, and long-term survival of new cells in the adult zebrafish brain. *J. Comp. Neurol.* *488*, 290–319.

DISCUSSION

Duplication de génome et expansion de la famille Sox

La famille des facteurs de transcription Sox joue un rôle crucial dans le développement et l'homéostasie des animaux. Cette famille, qui trouve son origine chez l'ancêtre commun des animaux, s'est agrandie par de multiples événements de duplication et divergence, et compte chez les mammifères 20 membres. Au cours des années 2000, des études ont montré que certains gènes *sox* des vertébrés sont dupliqués chez le poisson-zèbre et quelques autres espèces de poissons téléostéens, et ainsi suggéré pour la famille Sox des événements de duplication additionnels chez les poissons téléostéens. (De Martino *et al.* 2000, Chiang *et al.* 2001, Cresko *et al.* 2003, Kluever *et al.* 2005, Okuda *et al.* 2006, Braasch *et al.* 2009). L'étude présentée dans ce manuscrit récapitule le paysage des gènes *sox* dans dix génomes de poissons téléostéens, révèle une expansion de l'ensemble de la famille Sox dans ce vaste groupe de vertébrés et démontre que cette expansion est essentiellement due à la duplication de génome survenue à la base des poissons téléostéens. En outre, cette étude indique un taux de rétention des gènes *sox* dupliqués très élevé (55,5%) par rapport au taux de rétention moyen des gènes dupliqués dans les génomes des poissons téléostéens après la duplication de génome (12% à 24% ; Postlethwait *et al.* 2000, Jaillon *et al.* 2004, Woods *et al.* 2005, Brunet *et al.* 2006, Kassahn *et al.* 2009). Cependant, les gènes qui codent des facteurs de transcription et les gènes du développement sont des catégories fonctionnelles sur-représentées parmi les gènes retenus en deux copies après duplication de génome (Blanc & Wolfe 2004, Brunet *et al.* 2006, Aury *et al.* 2006, Kassahn *et al.* 2009). Ainsi, le fort taux de rétention des gènes *sox* en deux copies, qui codent des facteurs de transcription, est en accord avec ces études antérieures.

Par ailleurs, ce premier résultat révèle l'importance de la duplication de génome pour l'expansion de la famille des gènes *sox* chez les téléostéens. Nous remarquons en effet que les duplications segmentaires ont une contribution *quasi* nulle dans l'enrichissement en gènes *sox* des dix génomes téléostéens étudiés (observation sur la base des génomes et transcriptomes disponibles). En effet, sur une période évolutive d'environ 250 Ma (date estimée de la duplication de génome, Van de Peer *et al.* 2009), une unique duplication en tandem (du gène *sox17*) survenue après la duplication de génome chez l'ancêtre commun des téléostéens a été préservée dans les génomes téléostéens analysés. Ce constat ne signifie pas que les gènes *sox* ne sont *quasi* jamais

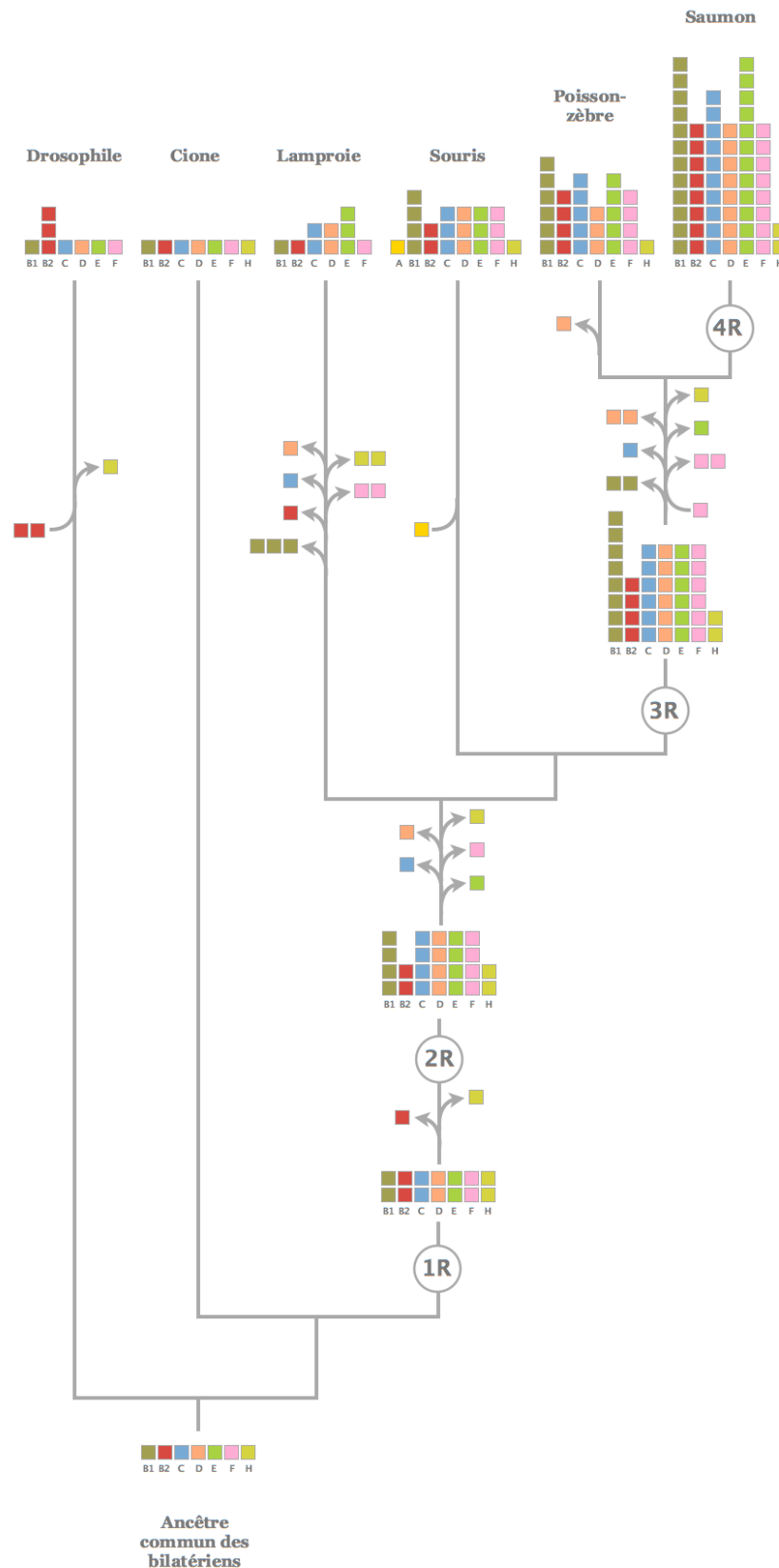


Figure 17: Scénario pour l'expansion de la famille Sox chez les bilatériens. Les duplications de génome survenues au cours de l'évolution des vertébrés sont indiquées (1R, 2R, 3R, et 4R). Les flèches tournées vers l'extérieur indiquent les événements de délétion. Les événements de duplication segmentaire et la divergence allélique du gène Sry sont représentés par une courbe qui rejoint la branche.

dupliqués par duplication segmentaire, mais qu'il y a éventuellement un effet délétère immédiat de la duplication segmentaire de ce type de gènes qui empêche la fixation de la nouvelle copie dans la population (Conant & Wolfe 2008). En effet, la duplication segmentaire d'un gène *sox*, au même titre que tous les autres gènes codant des protéines impliquées dans de nombreuses interactions, peut conduire éventuellement à des conflits de dosage chez les individus mutants, et en cela être délétère. Par conséquent, ces données suggèrent que la duplication de génome (qui *a priori* n'altère pas l'équilibre du dosage) survenue à la base des téléostéens a fourni une unique opportunité pour enrichir le répertoire de gènes *sox* dans les génomes téléostéens. De même, il est intéressant de remarquer que l'expansion de la famille des gènes *hox* chez les téléostéens, qui codent également des facteurs de transcription, est uniquement due à la duplication de génome survenue à la base des téléostéens (Prince & Pickett 2002).

Enfin, cette dernière hypothèse est en accord avec une expansion récurrente de la famille Sox observée au sein des animaux chordés dans les différents groupes qui possèdent un ancêtre commun polyploïde (les vertébrés, les téléostéens, et les salmonidés notamment). La figure 17 présente un scénario hypothétique qui récapitule quelques événements de duplication complète de la famille Sox probablement survenus au cours de l'évolution des vertébrés. L'ancêtre commun des bilatériens possédait probablement les 7 gènes *SoxB1*, *SoxB2*, *SoxC*, *SoxD*, *SoxE*, *SoxF* et *SoxH* qui sont pour la majorité présents chez l'ensemble des bilatériens actuels (les espèces choisies pour représenter les protostomiens (*Drosophila*) et les agnathes (*Lamproie*) ont chacune perdu le gène *SoxH*). Dans ce scénario, les deux duplications de génome à la base des vertébrés sont survenues avant la divergence des agnathes et des gnathostomes (représentés par la souris, le poisson-zèbre et le saumon), comme le suggèrent les derniers travaux de génomique comparative réalisés par Smith *et al.* (2013). Suite aux deux duplications de génome, et à quelques pertes subséquentes de gènes, le contenu en gènes Sox dans chacun des groupes nouvellement formés B1, B2, C, D, E, F et H chez l'ancêtre commun des vertébrés était probablement semblable à celui de la souris (le groupe A, qui contient le gène *Sry*, est spécifique des mammifères). Il est intéressant de remarquer que le positionnement des deux duplications de génome à la base de tous les vertébrés suppose en revanche de nombreuses pertes de gènes *sox* dans la lignée des agnathes. La duplication de génome à la base des téléostéens a conduit à l'enrichissement du répertoire de gènes *sox* spécifiquement dans ce grand groupe d'animaux vertébrés. Enfin, la duplication de génome additionnelle survenue à la

base des salmonidés est certainement à l'origine d'une nouvelle expansion de la famille Sox observée dans ce groupe. Ainsi, la figure 17 illustre notre hypothèse qui accorde une importance majeure aux duplications de génome dans l'expansion de la famille Sox au cours de l'évolution des vertébrés. Selon cette hypothèse, les duplications de génome peuvent conduire par l'enrichissement du répertoire de gènes du développement (par exemple les familles Hox et Sox) à une diversification phénotypique des espèces qui aurait eu peu de chance de survenir autrement (Freeling & Thomas 2006).

Enfin, cette étude confirme que le groupe des téléostéens constitue un modèle pertinent pour évaluer la contribution des duplications de génome dans la diversification du contenu génique (notamment des familles multigéniques) des génomes des animaux vertébrés, et étudier la portée évolutive des duplications de génome. Nous pouvons ajouter que le groupe des salmonidés, qui contient les espèces truite et saumon dont les génomes sont en cours d'annotation, et qui partagent un ancêtre polyploïde plus récent (~100 Ma), peut également constituer à l'avenir un modèle approprié pour étudier la dynamique évolutive des familles de gènes dans des génomes en cours de diploïdisation.

Devenir des gènes sox dupliqués chez les téléostéens

L'étude du devenir des gènes dupliqués après un événement de polyploïdisation, et notamment des mécanismes évolutifs qui ont conduit à la préservation sur le long terme des deux copies d'un gène, est également nécessaire pour comprendre l'impact des duplications de génome sur la diversification des familles de gènes, et l'évolution des génomes. Cette étude montre que la majorité (plus de 80%) des gènes *sox* préservés en deux copies dans les génomes téléostéens analysés évoluent de manière asymétrique en terme de dN/dS dans leur séquence codante. Typiquement, une copie est conservée et caractérisée par un dN/dS très faible et proche de celui du singleton orthologue des espèces non-téléostéens, tandis que l'autre copie est divergente et caractérisée par un dN/dS significativement plus élevé. L'augmentation du nombre de substitutions non-synonymes dans une des deux copies suggère des différences en terme de contrainte sélective entre les deux copies, avec probablement un relâchement de la sélection négative sur la copie divergente. Il est intéressant de noter que la copie divergente est toujours la même (orthologue) dans l'ensemble des espèces que nous avons analysées. En outre, la copie perdue dans certaines ou toutes les espèces est également toujours la même. Ces observations suggèrent que les deux copies pour chaque paire de *sox* étaient

déjà différentes en terme de contrainte sélective chez l'ancêtre commun des poissons téléostéens étudiés (qui correspond au dernier ancêtre commun des otocéphales et eutélostéens). Il est difficile d'estimer le temps qui sépare la duplication de génome de la divergence des deux groupes otocéphales et eutélostéens (du fait des incertitudes quant à la datation de ces deux événements : 226-350 Ma et 250-280 Ma respectivement ; Van de Peer *et al.* 2009, Near *et al.* 2012), mais nous pouvons supposer que les différences entre les deux paralogues chez l'ancêtre commun des deux groupes se sont accumulées en quelques dizaines de millions d'années. Ainsi, nous pouvons faire l'hypothèse que la destinée évolutive des paires de *sox* était en partie fixée chez l'ancêtre commun des espèces analysées, avec probablement une copie plus dispensable que l'autre. Quelques auteurs proposent que des différences dans l'environnement chromatinien des gènes dupliqués (modifications épigénétiques par exemple), acquises rapidement après la duplication de génome ou héritées des parents (si allopolyploïdisation), peuvent influencer le devenir des gènes dupliqués (Scannell *et al.* 2006, Thomas *et al.* 2006, Sémon & Wolfe 2007a).

La préservation sur le long terme des gènes dupliqués peut être expliquée par une évolution des deux copies par sous-fonctionnalisation ou néo-fonctionnalisation (Lynch & Conery 2000). Néanmoins, une fois préservées par sous-fonctionnalisation, des mutations qui attribuent une nouvelle fonction pour une copie (néo-fonctionnalisation) peuvent survenir sur de longues périodes évolutives (He & Zhang 2005, Rastogi & Liberles 2005, Conant & Wolfe 2008, Bridgham *et al.* 2008). Certains auteurs proposent que ces deux modèles de préservation peuvent être distingués sur la base des séquences codantes, car ils conduisent respectivement à une évolution symétrique et asymétrique des séquences codantes des deux paralogues en terme de dN/dS (Chain & Evans 2006, Sémon & Wolfe 2007a, Innan & Kondrashov 2010). En effet, les substitutions non-synonymes qui changent la séquence des acides aminés de la protéine sont plus affectées par la sélection naturelle que les substitutions synonymes. Ainsi, comme mentionné précédemment, le rapport dN/dS donne une indication des contraintes évolutives qui caractérisent chacun des deux paralogues. Le relâchement de la sélection négative et/ou une évolution adaptative sur les deux copies, attendu selon les modèles de sous-fonctionnalisation (*duplication, dégénération et complémentation* et *évitement du conflit adaptatif*), peuvent être ainsi mis en évidence par une augmentation symétrique du rapport dN/dS entre les deux copies. En revanche, le relâchement de la sélection négative et/ou une évolution adaptative d'une seule copie, attendu selon le

modèle de néo-fonctionnalisation, aboutira à une évolution asymétrique des deux paralogues en terme de dN/dS (Chain & Evans 2006). Selon cette hypothèse, l'évolution asymétrique de la majorité des *sox* dupliqués dans les génomes téléostéens suggèrerait une préservation par néo-fonctionnalisation d'une des deux copies dans certaines espèces. Néanmoins, les deux copies ont pu initialement être préservées par sous-fonctionnalisation (He & Zhang 2005, Rastogi & Liberles 2005). L'étude de la dynamique évolutive des séquences non-codantes pour quelques gènes *sox* dupliqués montre que les copies divergentes dans leur séquence codante tendent à être également divergentes dans leur séquence non-codante. Ainsi, les éléments très conservés des séquences non-codantes (qui sont de potentiels éléments cis-régulateurs) tendent à être distribués de manière asymétrique entre les deux paralogues. Nous avons suggéré précédemment que les différences entre les deux copies en terme de contrainte sélective étaient probablement présentes chez l'ancêtre commun des poissons analysés. Nous pouvons ajouter à notre hypothèse que l'accumulation de mutations dans la copie avec moins de contrainte sélective (au niveau des séquences codante et non-codante) au cours de l'évolution des poissons téléostéens peut avoir conduit à sa préservation par néo-fonctionnalisation ou spécialisation (pour une sous-fonction ancestrale) dans certaines espèces, et non-fonctionnalisation dans d'autres espèces. Après duplication de génome, l'évolution lignée-spécifique des paires de gènes sur de longue période évolutive peut avoir contribué à la diversification du contenu et des fonctions de la famille Sox chez les téléostéens.

Expansion de la famille Sox : diversification fonctionnelle ?

La duplication complète d'une famille de facteurs de transcription suite à un événement de polyploïdisation, peut être, plus qu'une opportunité pour la diversification fonctionnelle (par co-optation des facteurs dans des contextes cellulaires nouveaux), une opportunité pour soulager la charge de pléiotropie des différents membres de la famille, connus pour être impliqués dans de multiples processus physiologiques différents. Nous avons vu précédemment que les gènes *Sox* chez les vertébrés sont classés en 7 groupes de paralogues (le 8ème étant spécifique des mammifères). Selon les espèces vertébrés, chaque groupe contient plus ou moins de gènes paralogues selon que 2, 3 ou 4 événements de duplication de génome soient survenus au cours de leur histoire évolutive, et selon les plus ou moins nombreuses pertes de gènes subséquentes dans chaque groupe (figure 17). Souvent, plusieurs membres d'un même groupe de

paralogues sont exprimés dans un même tissu, et les gènes paralogues tendent à avoir en partie des fonctions redondantes (Lefebvre *et al.* 2007, Guth & Wegner 2008, Dy *et al.* 2008). Par exemple, les trois gènes *Sox4*, *Sox11* et *Sox12* qui constituent les paralogues du groupe C chez les vertébrés possèdent en partie des fonctions redondantes au cours du développement neural et mésenchymateux de la souris (Bhattaram *et al.* 2010). Ainsi, les souris mutantes *Sox12*^{-/-} sont viables et sans malformations apparentes, ce qui révèle que les fonctions du gène *Sox12* sont redondantes avec les deux autres *Sox4* et *Sox11*. Néanmoins, les gènes du groupe C, en particulier *Sox4* et *Sox11*, possèdent également des fonctions non-redondantes qui peuvent être la conséquence soit d'une redistribution des différentes sous-fonctions du gène *SoxC* de l'ancêtre des chordés entre paralogues, soit d'un recrutement opportuniste dans un nouveau contexte cellulaire (Schilham *et al.* 1996, Sock *et al.* 2004). Chez les téléostéens, les paralogues surnuméraires du groupe C produits par la duplication de génome conservent des profils d'expression au cours du développement similaires à ceux observés pour leurs orthologues tétrapodes (Mavropoulos *et al.* 2005, Rimini *et al.* 1999, De Martino *et al.* 2000). Ainsi, chez les vertébrés, à l'instar du groupe C, les différents groupes de paralogues de la famille *Sox* sont probablement impliqués dans des processus biologiques similaires où ils ont en partie des fonctions redondantes (Lefebvre *et al.* 2007, Guth & Wegner 2008).

Il semble nécessaire, selon cette hypothèse, d'étudier l'évolution fonctionnelle des différents gènes dupliqués, non pas indépendamment les uns des autres, mais par groupe de gènes dupliqués appartenant à un même groupe fonctionnel. En effet, des brassages de fonctions (ou réarrangements fonctionnels) peuvent survenir entre des gènes qui appartiennent à un même groupe fonctionnel (du fait de leur redondance fonctionnelle), et ainsi des gènes non-orthologues peuvent réaliser la même fonction dans différentes espèces, si tant est qu'ils appartiennent au même groupe, ou à la même famille de gènes (Prince & Pickett 2002). Ainsi, le brassage de fonctions au sein des familles multigéniques peut expliquer la réalisation d'une même fonction par deux gènes non-orthologues dans deux espèces différentes. Par exemple, chez le poisson-zèbre, un réarrangement fonctionnel a conduit le gène *hoxb1b* à réaliser la fonction effectuée par *Hoxa1* chez la souris. Les deux gènes appartiennent chacun au groupe 1 de la famille Hox mais ne sont pas orthologues (Prince & Pickett 2002). Le brassage de fonctions peut également expliquer la perte d'un gène dont la fonction est réalisée par un autre gène appartenant au même groupe.



Figure 18: Expression des gènes sox chez quatre vertébrés adultes. Etude de l'expression des gènes sox des groupes B2, C, E et F chez trois poissons téléostéens et chez la souris dans sept tissus adultes. Les gènes sox dupliqués chez les poissons téléostéens sont représentés lorsqu'ils sont exprimés par un carré de couleur (copie a en haut, copie b en bas) et le singleton orthologue de la souris par un rectangle coloré. Les gènes sox non exprimés sont représentés par un carré ou rectangle gris. Les gènes perdus après les différentes duplications de génome sont représentés en blanc.

La duplication de génome survenue à la base des téléostéens peut avoir contribué à soulager la charge pléiotropique des différents gènes de la famille *Sox* chez les poissons téléostéens. De plus, le partage des différentes sous-fonctions entre les gènes dupliqués d'un même groupe fonctionnel peut avoir impliqué des brassages de fonctions. La figure 18 récapitule les profils d'expression des gènes sox appartenant aux 4 groupes B2, C, E, et F dans sept tissus adultes des poisson-zèbre, médaka, platy et souris. Une interprétation de ces données d'expression par groupe de gènes sox indique globalement une conservation des profils d'expression entre les différentes espèces au sein d'un même groupe fonctionnel, et suggère également des cas de brassages de

fonctions espèce-spécifiques. Au sein du groupe B2, les deux paralogues *Sox14* et *Sox21* de la souris, et les quatre paralogues *sox14a*, *sox14b*, *sox21a* et *sox21b* des téléostéens (lorsqu'ils sont préservés) sont exprimés principalement dans le cerveau, ce qui suggère une fonction cérébrale ancestrale des *SoxB2*. Après duplication, la redondance fonctionnelle entre les paralogues d'un même groupe peut expliquer la perte du paraglogue *sox14b* chez le médaka, et du paraglogue *sox21a* chez les médaka et platy, dans la mesure où la fonction peut être réalisée par un autre *soxB2* chez ces deux poissons. Au moins un paraglogue du groupe C des différentes espèces analysées est exprimé dans les gonades, la rate, les yeux et le cerveau. La fonction potentielle du groupe C dans la rate semble réalisée par des gènes non-orthologues dans les différentes espèces : *Sox12* est exprimé dans la rate de la souris et du poisson-zèbre (avec *sox4b*), tandis que ce sont les gènes *sox4a* et *sox4b* qui sont exprimés dans la rate des médaka et platy (qui ont perdu le gène *sox12*). De même, l'expression gonadique des *SoxC* semble conservée chez les quatre espèces de vertébrés étudiées, et cependant réalisée par différents membres du groupe (non-orthologues) selon les espèces. Ces données suggèrent un brassage des fonctions entre les paralogues du groupe C au niveau des tissus adultes où ils sont majoritairement exprimés. En outre, les différents gènes d'un même groupe qui sont exprimés dans un même tissu peuvent à une autre échelle (cellulaire ou sub-cellulaire) réaliser différentes sous-fonctions. Notre étude par hybridation *in situ* de l'expression des deux gènes *sox11a* et *sox11b* (groupe C) du platy montre que le gène *sox11b* est exprimé uniquement dans des zones très restreintes du cerveau. Nous avons conclu à une potentielle néo-fonctionnalisation de ce gène dans le cerveau du platy. Néanmoins, ce résultat peut également supposer une spécialisation de ce gène dans une sous-fonction cérébrale qui peut être réalisée chez les autres espèces par un autre membre du groupe C. Au sein du groupe E, les différents paralogues des quatre espèces sont exprimés principalement dans les gonades, les yeux et le cerveau, en accord avec les rôles attribué aux *Sox8*, *Sox9* et *Sox10* dans ces tissus chez les vertébrés (Guth & Wegner 2008). Néanmoins, en particulier dans les gonades, les gènes exprimés dans les différents poissons téléostéens ne sont pas orthologues, ce qui suggère des réarrangements fonctionnels lignée-spécifiques. Par exemple, les gènes *sox8a* et *sox10b* sont exprimés dans les testicules du platy, tandis que le médaka exprime uniquement le gène *sox9b*. Au sein du groupe F, toutes les copies surnuméraires produites par la duplication de génome ont été perdues, probablement chez l'ancêtre commun des poissons analysés. Les données d'expression obtenues pour les différents paralogues du groupe F montrent une plus faible conservation des profils d'expression entre les

téléostéens et la souris. Ainsi, l'expression dans la rate, les muscles et dans les yeux des trois *sox7*, *sox17* et *sox18* est uniquement détectée chez les poissons téléostéens. En revanche, il est intéressant de remarquer que les gènes du groupe F sont exprimés chez les téléostéens de manière redondante dans la *quasi* totalité des tissus et dans les trois espèces de poissons téléostéens étudiées.

Un travail semblable réalisé sur un plus grand nombre de tissus (embryonnaires et adultes), par hybridation *in situ*, et sur plusieurs autres espèces de vertébrés, pourrait donner des indications sur l'évolution fonctionnelle de la famille Sox chez les vertébrés. Il est probable que cette étude indique une conservation des domaines d'expression au sein des groupes entre différentes espèces, suggérant une conservation des fonctions biologiques, avec pour chaque espèce la spécialisation fonctionnelle de certains gènes, la perte par non-fonctionnalisation de certains autres gènes, et l'illustration d'une évolution bricoleuse, contingente et opportuniste, suggérée il y a plusieurs décennies par François Jacob dans son livre «Le jeu des possibles» (Jacob, 1981).

Duplication de la famille Sox : impact sur la spéciation ?

La spéciation est la naissance à partir d'une seule espèce de deux espèces filles distinctes. L'étude de la spéciation consiste à comprendre les mécanismes d'isolement reproductif qui conduisent des populations à ne plus échanger de matériel génétique. Selon Werth et Windham (1991) puis Lynch et Force (2000), les duplications de gènes peuvent favoriser la séparation reproductrice post-zygotique entre deux populations, et par conséquent éventuellement la formation de nouvelles espèces. Ils postulent que la perte réciproque des gènes dupliqués ou une sous-fonctionnalisation différente des deux copies dans deux populations en allopatrie peut conduire à des incompatibilités génomiques chez les hybrides et avoir des conséquences sur leur viabilité (introduction, figure 14). L'importance de ces mécanismes génétiques est tout particulièrement reconnue après une duplication de génome (Lynch & Force 2000, Taylor *et al.* 2001, Postlethwait *et al.* 2004, Volff 2005, Scannell *et al.* 2006, Van de Peer *et al.* 2009). Notre étude du contenu en gènes *sox* dans dix génomes de poissons téléostéens, et de l'évolution des profils d'expression des gènes dupliqués dans différentes espèces de téléostéens, suggère une évolution lignée-spécifique de la famille Sox qui a pu contribuer à la divergence entre lignées et éventuellement aux différents événements de spéciation survenus depuis la duplication de génome.

Le devenir des gènes *sox* dupliqués, selon les événements de perte survenus après duplication de génome, peut aboutir à différentes classes de patrons de perte (figure 19). Après duplication, les deux copies peuvent avoir été préservées dans toutes les espèces analysées (aucune perte, classe I). C'est le cas des gènes *sox4* et *sox9* qui sont présents à l'état dupliqué dans les dix génomes de poissons téléostéens étudiés. Cependant, une des deux copies (toujours la même) peut avoir été perdue dans certaines espèces (classe II) ou dans toutes les espèces analysées (classe III). La classe II concerne les 8 gènes dupliqués *sox1*, *sox6*, *sox8*, *sox10*, *sox11*, *sox19*, *sox14*, et *sox21* dont une copie a été perdue de manière lignée-spécifique ou espèce-spécifique au cours de l'évolution des poissons téléostéens. La figure illustre le cas de *sox11* dont une même copie est perdue uniquement chez le médaka, le tétraodon, et le fugu. La classe III concerne les 7 gènes initialement dupliqués *sox2*, *sox3*, *sox5*, *sox7*, *sox13*, *sox17*, *sox18* à la base des téléostéens, dont une même copie a été perdue très certainement chez l'ancêtre commun des différentes espèces étudiées. Enfin, après duplication, les deux copies peuvent avoir été perdues au cours de l'évolution de certaines espèces et/ou lignées (classe IV). C'est le cas des gènes *sox12* et *sox30* dont les deux copies sont absentes des génomes de plusieurs espèces de poissons téléostéens analysés.

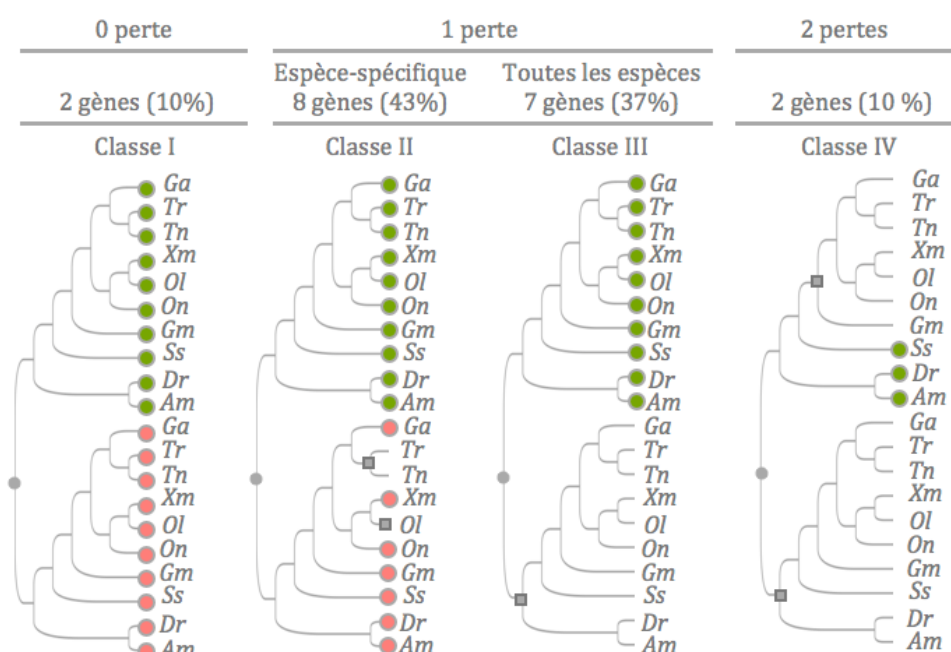


Figure 19: Patrons de perte des gènes *sox* après duplication. Les cercles rouges et verts représentent chacun un paraglogue issu de la duplication de génome. Le cercle gris représente l'événement de duplication de génome et les carrés gris indiquent les événements de perte les plus parcimonieux d'une copie après duplication. Les cas des gènes *sox11* et *sox12* illustrent respectivement une perte espèce-spécifique d'un même orthologue (classe II) et la perte des deux paralogues dans une même espèce (classe IV).

Nous remarquons que la majorité des gènes *sox* appartiennent à la classe II (43%). La présence des gènes *sox* en singleton dans certaines espèces et en deux copies dans certaines autres espèces (relation 2:1) peut avoir contribué à la divergence entre les génomes des poissons téléostéens. En effet, certains gamètes F_2 issus des hybrides produits par le croisement de deux populations qui ont une relation 2:1 pour différents loci seront dépourvus au moins d'une sous-fonction et, si elle est essentielle, non-viables (figure 20 ; Sémon & Wolfe 2007b). De la même manière, les gènes *sox* présents en singleton dans certaines espèces et absents dans certaines autres espèces (classe IV, relation 1:0) peuvent éventuellement avoir contribué à la divergence des génomes, et au développement des incompatibilités génomiques entre les espèces. Enfin, les deux gènes dupliqués dans l'ensemble des espèces analysées (classe I) peuvent avoir sous-fonctionnalisé différemment dans différentes lignées, et participé à la divergence des génomes téléostéens. En accord avec cette hypothèse, l'étude comparative des patrons d'expression de ces deux gènes réalisée à partir des trois poisson-zèbre, médaka et platy montre une divergence des patrons d'expression entre orthologues qui suggèrent des événements de sous-fonctionnalisation et/ou néo-fonctionnalisation lignée-spécifiques. Finalement, la duplication complète de la famille Sox chez les téléostéens, et son évolution (modulation du contenu et réarrangements fonctionnels) lignée-spécifique peut avoir contribué à la divergence entre lignées, et éventuellement aux différents événements de spéciation. Ces mécanismes génétiques, à l'échelle des centaines de gènes dupliqués au cours de la duplication de génome, peuvent ainsi avoir été impliqués dans la radiation évolutive des poissons téléostéens.

Enfin, l'étude du contenu de la famille Sox dans différents génomes de poissons téléostéens montre que des événements de non-fonctionnalisation sont survenus dans différentes lignées, pour certains probablement peu de temps après la duplication de génome, mais pour d'autres beaucoup plus récemment. Par exemple, il est très vraisemblable que la seconde copie des sept gènes *sox* en singleton dans toutes les espèces analysées ait été perdue peu de temps après la duplication de génome, probablement au cours du processus de diploïdisation (classe III). En revanche, les relations 2:1 pour certains gènes *sox* entre deux espèces apparentées suggèrent que des événements de non-fonctionnalisation sont survenus très longtemps après la duplication de génome (classe II). Par exemple, *sox14* est présent en deux copies chez le

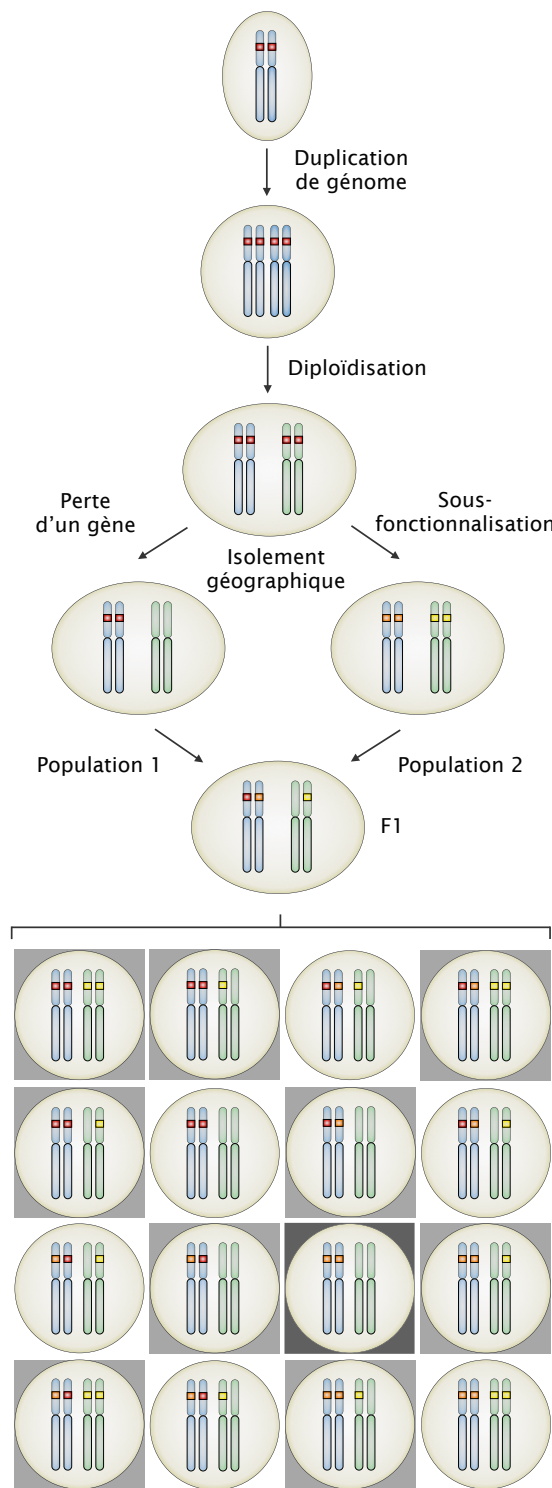


Figure 20: Non-fonctionnalisation lignée-spécifique après duplication. Les bandes rouges sur les chromosomes représentent un locus (ici un gène) dupliqué au cours d'un événement de polyploïdisation. Après diploïdisation, les deux copies du gène sont présentes sur deux chromosomes différents. Après isolement géographique, une population a perdu une copie, tandis que dans la seconde population, les deux copies ont évolué par sous-fonctionnalisation et/ou néo-fonctionnalisation (le schéma illustre un cas de sous-fonctionnalisation, bandes orange ou jaune sur les chromosomes). Après hybridation des deux populations, un-seizième des individus F2 issus des hybrides recevra un allèle non-fonctionnel et un allèle qui a perdu une sous-fonction, et sera probablement stérile et/ou non viable (carré gris foncé). D'autres individus F2 présenteront éventuellement des problèmes de conflit de dosage (carrés gris).

takifugu et en singleton chez le tétraodon, ce qui indique que l'ancêtre commun des deux poissons téléostéens, qui vivait à la toute fin du crétacé (~70 Ma), possédait deux copies du gène *sox14* (Near *et al.* 2012). Ces données montrent que les gènes dupliqués peuvent être préservés en deux copies dans les génomes téléostéens plus de 200 Ma après la duplication de génome, mais surtout que des événements de non-fonctionnalisation peuvent survenir après de très longues périodes évolutives (ici chez le tétraodon). De même, les premières indications fonctionnelles obtenues chez le poisson-zèbre et le platy suggèrent deux trajectoires évolutives différentes pour les copies *sox11* chez ces deux espèces empruntées après la divergence des deux lignées, plusieurs dizaines de millions d'années après la duplication de génome. Ces résultats supportent l'hypothèse que les gènes dupliqués, une fois préservés, peuvent emprunter dans certaines lignées de nouvelle trajectoire évolutive longtemps après duplication, et promouvoir la spéciation et la diversification phénotypique sur de longues périodes évolutives (Conant & Wolfe 2008, Van de Peer *et al.* 2009).

REFERENCES

- Abbasi, A.A. (2010). Piecemeal or big bangs: correlating the vertebrate evolution with proposed models of gene expansion events. *Nat. Rev. Genet.* **11**, 166.
- Abitua, P.B., Wagner, E., Navarrete, I.A., and Levine, M. (2012). Identification of a rudimentary neural crest in a non-vertebrate chordate. *Nature* **492**, 104–107.
- Ahn, D., You, K.-H., and Kim, C.-H. (2012). Evolution of the *tbx6/16* subfamily genes in vertebrates: insights from zebrafish. *Mol. Biol. Evol.* **29**, 3959–3983.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y., et al. (1998). Zebrafish *hox* clusters and vertebrate genome evolution. *SCIENCE* **282**, 1711–1714.
- Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310.
- Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178.
- Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., and Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.* **25**, 2453–2464.
- Bhattaram, P., Penzo-Mendez, A., Sock, E., Colmenares, C., Kaneko, K., Vassilev, A., DePamphilis, M., Wegner, M., and Lefebvre, V. (2010). Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nat. Commun.* **1**.
- Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.
- Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.
- Bowles, J., Schepers, G., and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol.* **227**, 239–255.
- Braasch, I., Brunet, F., Volff, J., and Schartl, M. (2009). Pigmentation Pathway Evolution after Whole-Genome Duplication in Fish. *GENOME Biol. Evol.* **1**, 479–493.
- Bridgham, J.T., Brown, J.E., Rodríguez-Marí, A., Catchen, J.M., and Thornton, J.W. (2008). Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors. *PLoS Genet* **4**, e1000191.
- Bronner, M.E., and LeDouarin, N.M. (2012). Development and evolution of the neural crest: An overview. *Dev. Biol.* **366**, 2–9.
- Brunet, F., Crollius, H., Paris, M., Aury, J., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23**, 1808–1816.
- Cañestro, C., Albalat, R., Irimia, M., and Garcia-Fernández, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Semin. Cell Dev. Biol.* **24**, 83–94.
- Chain, F.J.J., and Evans, B.J. (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genet.* **2**, e56.

- Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* *14*, 744–744.
- Chiang, E.F.-L., Yan, Y.-L., Guiguen, Y., Postlethwait, J., and Chung, B. (2001). Two Cyp19 (P450 Aromatase) Genes on Duplicated Zebrafish Chromosomes Are Expressed in Ovary or Brain. *Mol. Biol. Evol.* *18*, 542–550.
- Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* *6*, 836–846.
- Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* *9*, 938–950.
- Corbin, C.J., Mapes, S.M., Marcos, J., Shackleton, C.H., Morrow, D., Safe, S., Wise, T., Ford, J.J., and Conley, A.J. (2004). Paralogues of porcine aromatase cytochrome P450: a novel hydroxylase activity is associated with the survival of a duplicated gene. *Endocrinology* *145*, 2157–2164.
- Coriat, A.M., Müller, U., Harry, J.L., Uwanogho, D., and Sharpe, P.T. (1993). PCR amplification of SRY-related gene sequences reveals evolutionary conservation of the SRY-box motif. *Genome Res.* *2*, 218–222.
- Cresko, W., Yan, Y., Baltrus, D., Amores, A., Singer, A., Rodriguez-Mari, A., and Postlethwait, J. (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev. Dyn.* *228*, 480–489.
- Crow, K.D., and Wagner, G.P. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Mol. Biol. Evol.* *23*, 887–892.
- Dehal, P., and Boore, J. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLOS Biol.* *3*, 1700–1708.
- Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* *439*, 965–968.
- Denny, P., Swift, S., Connor, F., and Ashworth, A. (1992). An SRY-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein. *EMBO J.* *11*, 3705–3712.
- Donoghue, P., and Purnell, M. (2005). Genome duplication, extinction and vertebrate evolution. *TRENDS Ecol. Evol.* *20*, 312–319.
- Dy, P., Penzo-Méndez, A., Wang, H., Pedraza, C.E., Macklin, W.B., and Lefebvre, V. (2008). The three SoxC proteins—Sox4, Sox11 and Sox12—exhibit overlapping expression patterns and molecular properties. *Nucleic Acids Res.* *36*, 3101–3117.
- Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *GENETICS* *151*, 1531–1545.
- Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* *16*, 805–814.
- Gans, C., and Northcutt, R.G. (1983). Neural crest and the origin of vertebrates: a new head. *Science* *220*, 268–273.
- Gould, S.J., and Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. Ser. B Contain. Pap. Biol. Character R. Soc. G. B.* *205*, 581–598.
- Graves, J.A.M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell* *124*, 901–914.

- Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346**, 245–250.
- Guth, S., and Wegner, M. (2008). Having it both ways: Sox protein function between conservation and innovation. *Cell. Mol. LIFE Sci.* **65**, 3000–3018.
- He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157–1164.
- Hegarty, M., and Hiscock, S. (2007). Polyploidy: Doubling up for Evolutionary Success. *Curr. Biol.* **17**, R927–R929.
- Hoegg, S., and Meyer, A. (2005). Hox clusters as models for vertebrate genome evolution. *Trends Genet.* **21**, 421–424.
- Holland, L. (2009). Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nat. Rev. Neurosci.* **10**, 736–746.
- Holland, L.Z. (2013). Evolution of new characters after whole genome duplications: insights from amphioxus. *Semin. Cell Dev. Biol.* **24**, 101–109.
- Holland, N.D., and Chen, J. (2001). Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **23**, 142–151.
- Holland, P.W., Garcia-Fernández, J., Williams, N.A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Dev. Camb. Engl. Suppl.* 125–133.
- Van Hoof, A. (2005). Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171**, 1455–1461.
- Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256**, 119–124.
- Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108.
- Jacob, F. (1981). *Le jeu des possibles. Essai sur la diversité du vivant.* Editions Fayard.
- Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *NATURE* **431**, 946–957.
- Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467.
- Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
- Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. (2001). Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519.
- Kaessmann, H. (2009). More Than Just a Copy. *Science* **325**, 958–959.
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *GENOME Res.* **20**, 1313–1326.

- Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., and Ishibashi, T. (1996). Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci. U. S. A.* 93, 9096–9101.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719.
- Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C., and Ragan, M.A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19, 1404–1418.
- Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624.
- King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., et al. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451, 783–788.
- Klüber, N., Kondo, M., Herpin, A., Mitani, H., and Schartl, M. (2005). Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. *Dev. GENES Evol.* 215, 297–305.
- Kopelman, N.M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat. Genet.* 37, 588–589.
- Kuraku, S., Meyer, A., and Kuratani, S. (2009). Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.* 26, 47–59.
- Larroux, C., Fahey, B., Liubicich, D., Hinman, V.F., Gauthier, M., Gongora, M., Green, K., Worheide, G., Leys, S.P., and Degnan, B.M. (2006). Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol. Dev.* 8, 150–173.
- Larroux, C., Luke, G.N., Koopman, P., Rokhsar, D.S., Shimeld, S.M., and Degnan, B.M. (2008). Genesis and Expansion of Metazoan Transcription Factor Gene Classes. *Mol. Biol. Evol.* 25, 980–996.
- Lecointre, G., Gallut, C., Chanet, B., and Dettai, A. (2010). Du rififi chez les poissons. *Pour la Science* n 390.
- Lecointre, G., and Le Guyader, H. (2013). Classification phylogénétique du vivant - Tome 2. Editions Belin.
- Lefebvre, V., Dumitriu, B., Penzo-Mendez, A., Han, Y., and Pallavi, B. (2007). Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Int. J. Biochem. CELL Biol.* 39, 2195–2214.
- Lundin, L.G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16, 1–19.
- Lynch, M., and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *SCIENCE* 290, 1151–1155.
- Lynch, M., and Force, A.G. (2000). The origin of interspecific genomic incompatibility via gene duplication. *Am. Nat.* 156, 590–605.
- Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459.

- Makalowski, W. (2001). Are we polyploids? A brief history of one hypothesis. *Genome Res.* *11*, 667–670.
- Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci.* *107*, 9270–9274.
- Des Marais, D.L., and Rausher, M.D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* *454*, 762–765.
- De Martino, S., Yan, Y., Jowett, T., Postlethwait, J., Varga, Z., Ashworth, A., and Austin, C. (2000). Expression of *sox11* gene duplicates in zebrafish suggests the reciprocal loss of ancestral gene expression patterns in development. *Dev. Dyn.* *217*, 279–292.
- Mavropoulos, A., Devos, N., Biemar, F., Zecchin, E., Argenton, F., Edlund, H., Motte, P., Martial, J., and Peers, B. (2005). *sox4b* is a key player of pancreatic alpha cell differentiation in zebrafish. *Dev. Biol.* *285*, 211–223.
- Mazet, F., Hutt, J.A., Milloz, J., Millard, J., Graham, A., and Shimeld, S.M. (2005). Molecular evidence from *Ciona intestinalis* for the evolutionary origin of vertebrate sensory placodes. *Dev. Biol.* *282*, 494–508.
- Meyer, A., and Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* *11*, 699–704.
- Mu, L., Berti, L., Masserdotti, G., Covic, M., Michaelidis, T.M., Doberauer, K., Merz, K., Rehfeld, F., Haslinger, A., Wegner, M., et al. (2012). SoxC Transcription Factors Are Required for Neuronal Differentiation in Adult Hippocampal Neurogenesis. *J. Neurosci.* *32*, 3067–3080.
- Nanda, S., DeFalco, T.J., Hui Yong Loh, S., Phochanukul, N., Camara, N., Van Doren, M., and Russell, S. (2009). Sox100B, a Drosophila Group E Sox-domain Gene, Is Required for Somatic Testis Differentiation. *Sex. Dev.* *3*, 26–37.
- Naruse, K., Fukamachi, S., Mitani, H., Kondo, M., Matsuoka, T., Kondo, S., Hanamura, N., Morita, Y., Hasegawa, K., Nishigaki, R., et al. (2000). A detailed linkage map of Medaka, *Oryzias latipes*: Comparative genomics and genome evolution. *GENETICS* *154*, 1773–1784.
- Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J., and Mitani, H. (2004). A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* *14*, 820–828.
- Navarro-Martín, L., Galay-Burgos, M., Sweeney, G., and Piferrer, F. (2009). Different *sox17* transcripts during sex differentiation in sea bass, *Dicentrarchus labrax*. *Mol. Cell. Endocrinol.* *299*, 240–251.
- Navarro-Martín, L., Galay-Burgos, M., Piferrer, F., and Sweeney, G. (2012). Characterisation and expression during sex differentiation of *Sox19* from the sea bass *Dicentrarchus labrax*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* *163*, 316–323.
- Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M., and Smith, W.L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 13698–13703.
- Ner, S.S. (1992). HMGs everywhere. *Curr. Biol. CB* *2*, 208–210.
- Ohno, S. (1970). Evolution by gene duplication. Editions Springer-Verlag.
- Okuda, Y., Yoda, H., Uchikawa, M., Furutani-Seiki, M., Takeda, H., Kondoh, H., and Kamachi, Y. (2006). Comparative genomic and expression analysis of group B1 *sox* genes in zebrafish indicates their diversification during vertebrate evolution. *Dev. Dyn.* *235*, 811–825.

- Opazo, J.C., Butts, G.T., Nery, M.F., Storz, J.F., and Hoffmann, F.G. (2013). Whole-Genome Duplication and the Functional Diversification of Teleost Fish Hemoglobins. *Mol. Biol. Evol.* **30**, 140–153.
- Orengo, C.A., and Thornton, J.M. (2005). Protein families and their evolution-a structural perspective. *Annu. Rev. Biochem.* **74**, 867–900.
- Otto, S.P. (2007). The evolutionary consequences of polyploidy. *Cell* **131**, 452–462.
- Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197.
- Paps, J., Holland, P.W.H., and Shimeld, S.M. (2012). A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Brief. Funct. Genomics* **11**, 177–186.
- Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahoul, A., Cargill, M., Jones, P.G., et al. (2009). An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**, 995–998.
- Van de Peer, Y. (2004). Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* **5**, 752–763.
- Van de Peer, Y., Maere, S., and Meyer, A. (2009). OPINION The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732.
- Van de Peer, Y., Maere, S., and Meyer, A. (2010). 2R or not 2R is not the question anymore. *Nat. Rev. Genet.* **11**, 166.
- Phochanukul, N., and Russell, S. (2010). No backbone but lots of Sox: Invertebrate Sox genes. *Int. J. Biochem. CELL Biol.* **42**, 453–464.
- Piatigorsky, J., O'Brien, W.E., Norman, B.L., Kalumuck, K., Wistow, G.J., Borrás, T., Nickerson, J.M., and Wawrousek, E.F. (1988). Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 3479–3483.
- Postlethwait, J., Amores, A., Cresko, W., Singer, A., and Yan, Y.-L. (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet.* **20**, 481–490.
- Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**, 827–837.
- Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071.
- Ramsey, J., and Schemske, D.W. (1998). Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants. *Annu. Rev. Ecol. Syst.* **29**, 467–501.
- Rastogi, S., and Liberles, D.A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **5**, 28.
- Rimini, R., Beltrame, M., Argenton, F., Szymczak, D., Cotelli, F., and Bianchi, M. (1999). Expression patterns of zebrafish *sox11A*, *sox11B* and *sox21*. *Mech. Dev.* **89**, 167–171.
- Santos, E.M., Workman, V.L., Paull, G.C., Filby, A.L., Look, K.J.W.V., Kille, P., and Tyler, C.R. (2007). Molecular basis of sex and reproductive status in breeding zebrafish. *Physiol. Genomics* **30**, 111–122.

- Sarkar, A., and Hochedlinger, K. (2013). The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell* 12, 15–30.
- Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., and Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440, 341–345.
- Schartl, M. (2004). Sex chromosome evolution in non-mammalian vertebrates. *Curr. Opin. Genet. Dev.* 14, 634–641.
- Schilham, M.W., Oosterwegel, M.A., Moerer, P., Ya, J., de Boer, P.A., van de Wetering, M., Verbeek, S., Lamers, W.H., Kruisbeek, A.M., Cumano, A., et al. (1996). Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. *Nature* 380, 711–714.
- Sekido, R., and Lovell-Badge, R. (2008). Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. *Nature* 453, 930–934.
- Semon, M., and Wolfe, K. (2007a). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512.
- Semon, M., and Wolfe, K.H. (2007b). Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* 23, 108–112.
- Sémon, M., and Wolfe, K.H. (2007). Rearrangement Rate following the Whole-Genome Duplication in Teleosts. *Mol. Biol. Evol.* 24, 860–867.
- Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischau, A.M., Lovell-Badge, R., and Goodfellow, P.N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* 346, 240–244.
- Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M.S., Yandell, M.D., Manousaki, T., Meyer, A., Bloom, O.E., et al. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* 45, 415–421.
- Sock, E., Rettig, S.D., Enderich, J., Bösl, M.R., Tamm, E.R., and Wegner, M. (2004). Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Mol. Cell. Biol.* 24, 6635–6644.
- Stemple, D.L. (2005). Structure and function of the notochord: an essential organ for chordate. *Development* 132, 2503–2512.
- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and Collinearity in Plant Genomes. *Science* 320, 486–488.
- Taylor, J., and Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.* 38, 615–643.
- Taylor, J.S., Van de Peer, Y., and Meyer, A. (2001). Genome duplication, divergent resolution and speciation. *Trends Genet.* 17, 299–301.
- Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946.
- Thomson, J.M., Gaucher, E.A., Burgan, M.F., De Kee, D.W., Li, T., Aris, J.P., and Benner, S.A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat. Genet.* 37, 630–635.
- Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263.

- Volff, J. (2005). Genome evolution and biodiversity in teleost fish. *HEREDITY* 94, 280–294.
- Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54–61.
- Wegner, M. (1999). From head to toes: the multiple facets of Sox proteins. *Nucleic Acids Res.* 27, 1409–1420.
- Wegner, M. (2011). SOX after SOX: SOXession regulates neurogenesis. *Genes Dev.* 25, 2423–2428.
- Werth, C., and Windham, M. (1991). A Model for Divergent, Allopatric Speciation of Polyploid Pteridophytes Resulting from Silencing of Duplicate-Gene Expression. *Am. Nat.* 137, 515–526.
- Wistow, G., and Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Science* 236, 1554–1556.
- Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341.
- Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.
- Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. (2005). The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* 15, 1307–1314.
- Zapater, C., Chauvigné, F., Tingaud-Sequeira, A., Finn, R.N., and Cerdà, J. (2013). Primary oocyte transcriptional activation of *aqp1ab* by the nuclear progesterone receptor determines the pelagic egg phenotype of marine teleosts. *Dev. Biol.* 377, 345–362.

REMERCIEMENTS

Je voudrais tout d'abord m'adresser aux membres de mon jury Didier Casane, Marc Robinson-Rechavi, Marie Sémon, Élisabeth Huguet, et Jean-Nicolas Volff et les remercier d'avoir accepté de lire et évaluer ce travail.

Je tiens ensuite à remercier très sincèrement ma directrice de thèse Delphine Galiana pour la bienveillance et la confiance dont elle a fait preuve à mon égard. Je la remercie de m'avoir donné l'opportunité de développer ce projet de recherche qui rassemble des thématiques qui me tiennent à cœur, et de m'avoir accompagné tout au long de ces trois années de thèse.

Je tiens également à remercier l'ensemble de l'équipe « Génomique évolutive des poissons » avec qui ce fut un grand plaisir de travailler. Je remercie Jean-Nicolas Volff pour m'avoir accueilli dans son laboratoire, pour sa disponibilité, et pour ses discussions stimulantes. Je remercie Frédéric Brunet, avec qui le projet de thèse fut initié et développé au sein de l'équipe, ainsi que Magali Naville pour sa contribution au projet et pour son humour toujours bienvenu. Je souhaite remercier chaleureusement Domitille Chalopin pour sa gentillesse et je retiendrai la complicité discrète mais sincère que nous avons entretenue ces trois dernières années. Je remercie enfin mon stagiaire Luis Anthony Leon Florian pour son travail et sa bonne humeur apportés au cours de ma dernière année de thèse.

La réalisation expérimentale de ma thèse a été rendue possible grâce à la contribution de plusieurs personnes auxquelles ces quelques mots de remerciements sont tout spécialement dédiés. Je remercie en premier lieu Pierre Godement pour m'avoir initié avec succès aux expériences d'hybridation *in situ*, et pour ses nombreuses discussions sur la biologie du cerveau. Je remercie également Marie Sémon pour ses remarques pertinentes sur des questions bioinformatiques. Je dois évidemment remercier les personnes qui ont rendu possible mon travail sur les poissons : Laure Bernard, Bernard Perret et Robert Renard. Je pense également aux personnes qui m'ont permis de travailler sur le modèle murin : Denise Aubert, Marie Teixeira et Romain Guyot. Enfin,

j'adresse mes remerciements aux deux membres de mon comité de thèse Hugues Roest Crollius et Florence Ruggiero.

Finally, je remercie l'Institut de Génomique Fonctionnelle de Lyon, et notamment Vincent Laudet pour m'y avoir accueilli. Un merci tout particulier à l'ensemble des membres de cet institut qui ont contribué au plaisir que j'ai eu à venir travailler chaque jour. Je pense aux nombreuses discussions que j'ai pu avoir à l'heure du déjeuner avec Bachar Cheaib, Li Ling et François Bonneton. Je pense à la sympathie très bienvenue que m'ont portée Juliana Gutierrez Mazariegos et Mélisandre Téfit au cours des derniers mois de ma thèse. Je pense également à Helder Gomes Rodrigues, Nathalie Demoncheaux, Julie Carnesecchi, Damien Curton, Steven Nindorera-Badara et Kélig Pernelle que j'ai eu le plaisir de côtoyer au cours de ma thèse. Je pense à nos discussions musicales trop rares mais enrichissantes avec Sophie Pantalacci et Jean-Marc Vanacker. Je remercie l'équipe administrative, et les membres des services généraux, pour leur efficacité et leur gentillesse.

Je remercie enfin tout particulièrement François et Dane, avec qui j'ai pu vivre pleinement Lyon et rencontrer des personnes qui me sont devenues chères. Je n'oublie pas les amis de longue et de moins longue date, ainsi que mes proches, pour leur soutien sans faille.

ANNEXE

THE SOX FAMILY: EVOLUTION, BIOLOGY AND ITS ROLE IN DISEASES

The SRY-related HMG box, or Sox, gene family is named after the Sex-determining region of Y chromosome (SRY) gene discovery (Sinclair *et al.* 1990; Gubbay *et al.* 1990). This gene is present on the Y chromosome short arm in mammals and encodes a transcription factor with a characteristic DNA binding domain called HMG box domain. The SRY factor initiates male sex determination during mammalian development, and mutations in *SRY* gene cause sex-reversal (Hawkins 1993). The Sox family is an ancient transcription factors family, which originated in unicellular eukaryotes and acquired fundamental functions in the biology of animals (King *et al.* 2008). Sox factors are key players in animal development and adult homeostasis, from zygote totipotency acquisition to beta cells differentiation in vertebrate pancreas, from adult neurogenesis to sex determination (Sarkar & Hochedlinger 2013). A serial process of gene duplication and divergence during animal evolution has generated this huge functional diversity. Some Sox factors have conserved functions, such as in gonadal differentiation from fly to vertebrates, whereas others evolved new functions, such as in neural crest development in vertebrates. The importance of the Sox family in biological processes ensues roles in several human diseases, from genetic disorders to cancers. In this review, we first attempted to recapitulate the evolutionary history of the Sox family in animals, shaped by notable events such as whole genome duplication and gene losses, leading to the twenty Sox members in mammals. Then, we attempted to depict the spectacular functional diversification of the Sox family, in terms of molecular properties and physiological roles during development and adulthood, with an emphasis on vertebrates. Since its discovery, the Sox family has been described in a wide range of biological processes, and this review focuses on some of them, to illustrate the importance of Sox factors in animal biology and evolution. We finally reviewed genetic disorders and cancers in which Sox factors have critical roles. This review highlights the great interest of the Sox family in several biological fields, from evolutionary and developmental biology to molecular biology and health biology.

Discovery

The unifying feature of Sox proteins is the SRY-related high-mobility-group (HMG) domain which consists of 79 residues highly conserved across all members of the Sox family. Simultaneous discoveries in human and mouse of the sex-determining gene

(*SRY*) led to the identification of the first member of the Sox family in 1990 (Sinclair *et al.* 1990; Gubbay *et al.* 1990). The novel *SRY* gene showed similarities with chromosomal proteins HMG1 and HMG2 and therefore was classed among the HMG box superfamily (Ner 1992). Moreover, specific features in the *SRY*-HMG box domain led to the rapid identification of *SRY*-related HMG elsewhere in the genome and in different invertebrate and vertebrate species (Gubbay *et al.* 1990, Denny *et al.* 1992, Coriat *et al.* 1993). *SRY*-related genes have then been named *Sox* genes for *SRY*-related HMG box. Architecture and dynamic of the HMG domain were solved for several Sox proteins : Sry (Werner *et al.* 1995), Sox2 (Remenyi *et al.* 2003), Sox4 (van Houte *et al.* 1995, Jauch *et al.* 2012), Sox5 (Cary *et al.* 2001), and Sox17 (Palasingam *et al.* 2009, Gao *et al.* 2013). The prototypical HMG domain consists of three helices forming an L-shaped fold. The major wing encompasses the two first helices, and the third helices with the N-terminal extension form the minor wing (Crane-Robinson *et al.* 1998). Residues extending from the N-terminal half of the HMG domain (two first helices) mediate the majority of contacts with the minor groove side of DNA (Palasingam *et al.* 2009). Apart from the HMG domain, transactivation domains (TAD), phosphorylation and acetylation sites, and nuclear localization signals domains are characteristic features of Sox factors. These characteristics underlie the numerous properties of Sox factors: transcription factors, architectural proteins, interaction with partners, and shuttling proteins (Lefebvre *et al.* 2007, Malki *et al.* 2010).

Phylogeny and classification

Phylogenetic inferences on the basis of the HMG domain relate the relationship between dozens of *Sox* genes identified in vertebrate and invertebrate species. Thus, mammals have twenty *Sox* genes divided into eight monophyletic groups: the seven bilaterian-common groups B1, B2, C, D, E, F, and H, and the mammal-specific group A (Bowles *et al.* 2000, Larroux *et al.* 2008, Han *et al.* 2010, this review). Moreover, invertebrate species (protostomes and invertebrate deuterostomes) generally possess a single *Sox* gene for each group (Phochanukul & Russell 2010, Paps *et al.* 2012) whereas vertebrate genomes contain several *Sox* genes per group due to vertebrate-specific duplication events (discussed in detail below). As an example, the three *Sox5*, *Sox6* and *Sox13* members found in vertebrates are co-orthologues to the invertebrates *SoxD* (named *Sox102F* in *Drosophila melanogaster* and *SoxD* in *Ciona intestinalis*). Accumulating data from vertebrates genomes precise classification of the different Sox members. Thus, phylogenetic reconstruction show that the two *Sox15* and *Sox19* members previously

separated into two distinct groups according to their sequence disparity (Bowles *et al.* 2000) form one monophyletic group among *SoxB1* (figure 1). Syntenic analysis confirms the common origin of the two *Sox15* and *Sox19* in vertebrates (Okuda *et al.* 2006). Moreover, the group A, corresponding to the *Sry* gene in mammals, is an allelic variant of the ancient gene *Sox3* evolved during Y chromosome degradation (Graves *et al.* 2006). This group is therefore included in the *SoxB1* monophyletic group in the phylogeny and closed to the *Sox3* members. Finally, *SoxH* was recently found in several species, from acorn worm to mammals, suggesting its presence in stem bilaterians (Han *et al.* 2010). Interestingly, this group corresponds to a single gene in the majority of vertebrate and invertebrate species, and therefore seems to be duplication-resistant.

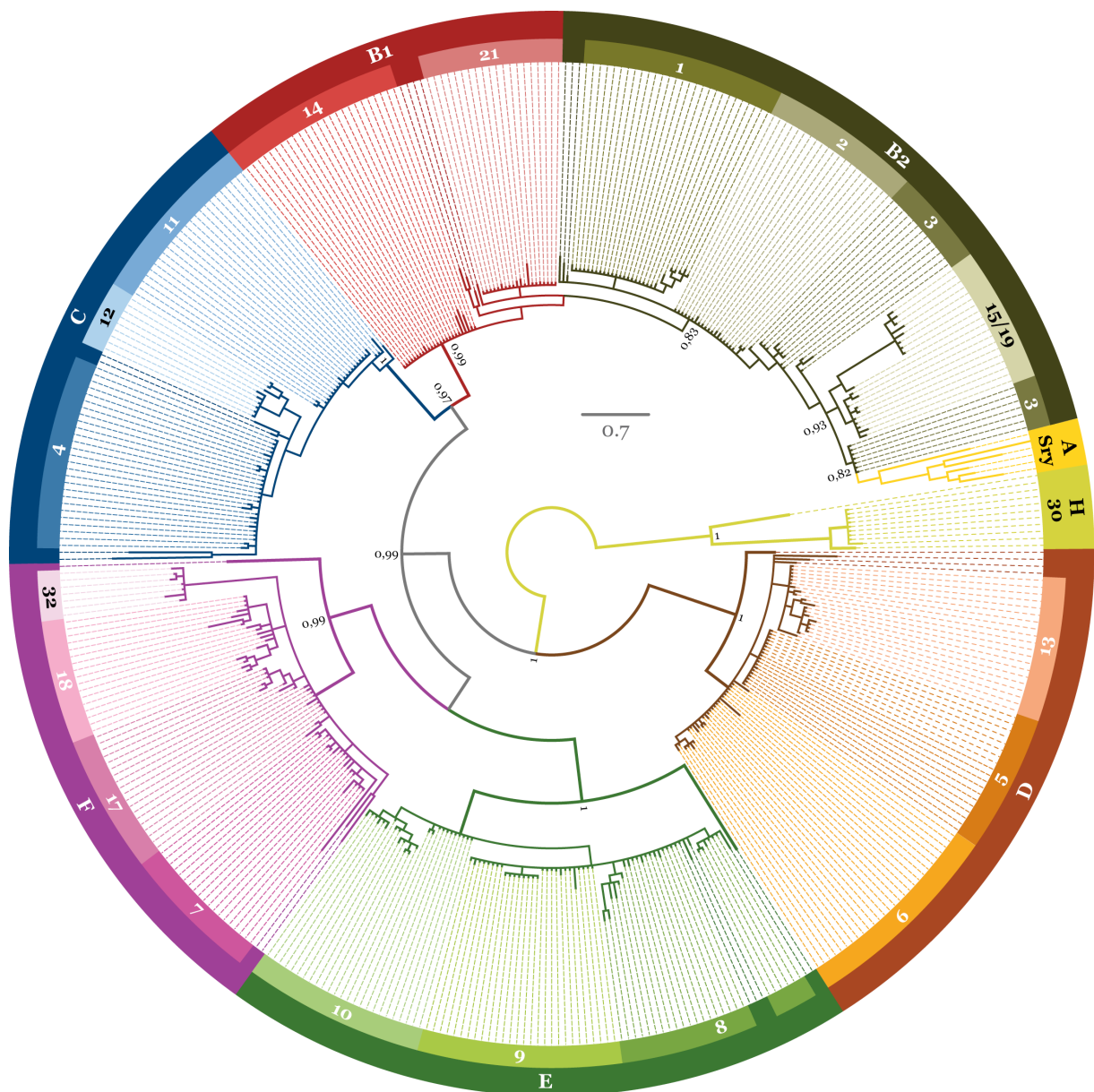


Figure 1. Evolutionary tree of vertebrate SRY-related HMG box (*Sox*) gene family. The Bayesian phylogeny is based on the protein sequence of the HMG domain.

Origin and expansion

The Sox family is an ancient class of transcription factors. Comparative genomics including poriferans, cniderians and bilaterians provide evidence for origin of the Sox genes family during the early evolution of metazoans (Larroux *et al.* 2006; Larroux *et al.* 2008). Moreover, the two Sox-like genes in the unicellular choanoflagellate suggest origin in unicellular eukaryotes (King *et al.* 2008). It is likely that gene and/or domain duplication events led to the seven *Sox* genes B1, B2, C, D, E, F and H in the bilaterians ancestor, which has been maintained until the present day in bilaterian species (Bowles *et al.* 2000, Jager *et al.* 2006). Invertebrate genomes generally contain a single *Sox* gene for each group, with occasional additional copies (Phochanukul & Russell 2010, Paps *et al.* 2012). During chordate evolution, the two round of whole genome duplication occurred near the base of vertebrates led to the expansion of many transcription factors families in vertebrates (Hokamp *et al.* 2003; Holland 2009). These notable events could explain both expansion and genomic organization of the Sox family in vertebrates (figure 2). Indeed, vertebrate genomes possess several *Sox* genes for each group dispersed in the genome. For instance, the four *SOX1*, *SOX2*, *SOX3* and *SOX15/19* are co-orthologues to the invertebrates *SoxB1* and located respectively on the four human chromosomes 13, 3, X and 17. Moreover, chromosomal segments containing these four human *SOX* genes show significant conservation of local synteny (Okuda *et al.* 2006). Whole genome duplication and subsequent inter-chromosomal rearrangements during rediploidization seem to be the most parsimonious explanation for *Sox* gene expansion and distribution in vertebrate genomes. Moreover, the fish-specific whole genome duplication led to the duplication of many vertebrate *Sox* genes in fishes (Voldoire *et al.* in preparation). As a result, the zebrafish, as well as many other fish genomes, contain more than twenty *sox* genes. Finally, salmonids have undergone an additional whole genome duplication and therefore possess around sixty Sox members. These observations show that the Sox family expansion seems to be mainly mediated by whole genome duplications during the evolution of vertebrates (figure 2). This idea implies that Sox family expansion and diversification would have been unlikely to arise otherwise than by whole genome duplication. The well-characterized Sox family, in terms of evolution and function, could constitute a relevant framework to study the relationship between biochemical features, gene duplicability and evolutionary constraints.

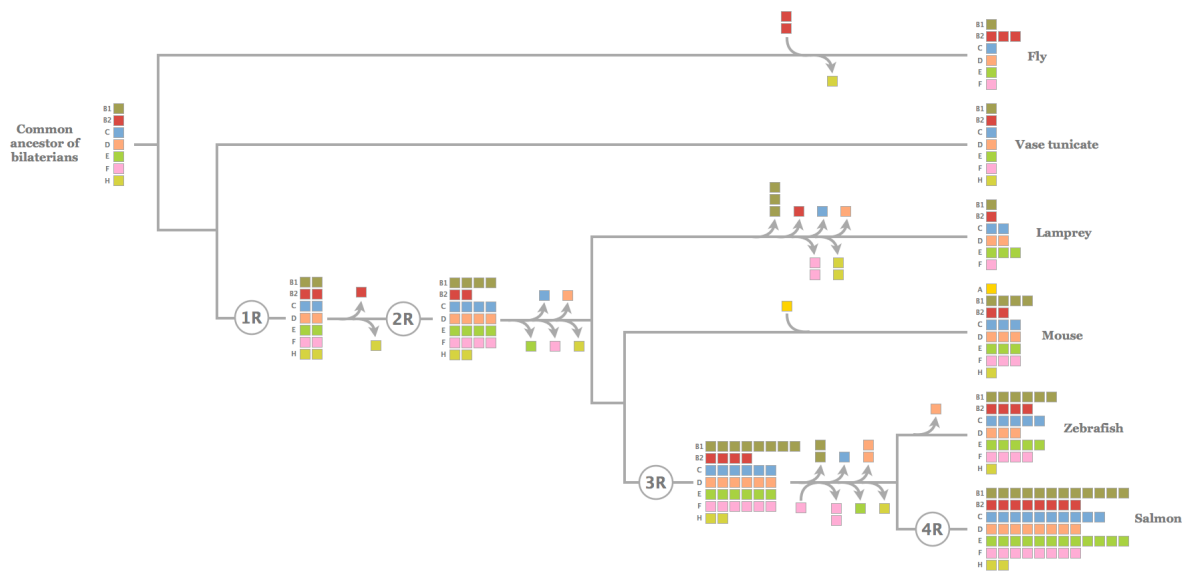


Figure 2. *Scenario for Sox evolution in bilaterians.* Vertebrate-specific whole genome duplication events are indicated (1R, 2R, 3R and 4R).

Diversification in function

At the molecular and cellular levels, the Sox factors possess the conserved function of stem and progenitor cell fate regulators. As an example, the well-characterized Sox2 members (group B1) is fundamental for the maintenance of neuronal stem cells (NSCs) during vertebrate neurodevelopment. Similarly, the Sox17 factor (Group F) is required for the maintenance of fetal and neonatal hematopoiesis stem cells in vertebrates (Sarkar & Hochedlinger 2013). However, at the organism level, the Sox members have evolved a wide range of distinct biological roles during embryogenesis. This functional diversity was generated partially through expression pattern divergence among Sox members after duplication events, mechanism theorized by the well-known subfunctionalization and neofunctionalization models (for a detailed review, see Conant & Wolfe 2008). For example, the three vertebrate *Sox8*, *Sox9* and *Sox10* have evolved new spatio-temporal expressions in vertebrate neural crest cells (Guth & Wegner 2008) (discussed below). Diversification in partner interactions constitutes a second way for functional diversification of the Sox members. However, overlapping and redundant functions among Sox members also characterized Sox evolution, exemplified by the three *Sox1*, *Sox2* and *Sox3* during neurogenesis (discussed in detail below). Finally, it worth noting that almost all Sox factors have several expression domains, at different time during development, and interact with several co-factors. This provides a cell-context dependent function to each Sox protein that enriched the size and breadth of function of the Sox family. In the following sections, we will present important

properties of Sox factors in biological processes, which also provide examples of their spectacular functional diversity, with an emphasis on vertebrates.

Sox factors in animals: from embryo to adulthood

Sox in neurodevelopment: transcription and pioneer factors

In vertebrates, the central nervous system develops from neural stem cells, which turn into neural precursor cells (NPCs). During embryogenesis, the neural plate folds into the neural tube where NPCs still proliferate before committing into the neuronal or glial lineages. Different Sox proteins act successively during neurogenesis. Sox2 present in neural stem cells is joined by the two Sox1 and Sox3 in NPCs. Then, these three SoxB1 are relayed by Sox4 and Sox11 (SoxC) during neuronal differentiation (Guth & Wegner 2008). The discovery that the single SoxB1 orthologue in the basal chordate amphioxus participates to the neural induction suggests an ancient and conserved role of Sox proteins in neurogenesis (Meulemans & Bronner-Fraser 2007). During neurodevelopment, these Sox members work as transcriptional activators and up-regulate hundreds of target genes (Wegner 2011). This function relies on both their DNA-binding properties and their *trans*-activation domain and requires co-activators recruitment. Interestingly, Sox proteins act not only as stage-specific transcription factors but also as pioneer factors. Architectural properties allow Sox proteins to keep chromatin near targeted enhancer in a poised state for the subsequent binding of other transcription factors. An elegant mechanism implies these Sox properties during neurodevelopment (Bergsland *et al.* 2011). Indeed, SoxB1 transcription factors activate their target genes in NCPs but also function as pioneer factors occupying neuron-specific enhancers. Thereafter, they are replaced by SoxC transcription factors once cells commit to the neuronal lineage. The function of pioneer factors of Sox members is also described during B-cell development, in which Sox2 prebinds enhancers later activated by Sox4 in differentiating B cells (Liber *et al.* 2010), suggesting a widespread pioneer function of Sox proteins during organ development. Pioneer factors emerge as important players during animal development, homeostasis as well as tumorigenesis (Magnani *et al.* 2011). The Sox family provides a relevant framework to deepen our understanding of the mode of action of pioneer factors during these biological processes.

Sox in adult homeostasis

Continual cell replacement from adult stem cells for the maintenance of adult tissues is called tissue homeostasis. Accumulating evidences indicate that Sox proteins role in cell fate decisions during embryogenesis continues into limited adult tissues (Sarkar & Hochedlinger 2013). In mammals, the hippocampal dentate gyrus and the olfactory bulb are the two regions in the brain where new neurons are generated throughout adulthood (Zhao *et al.* 2008). Sox2 maintain multipotency of adult neural stem cells (NSCs) in concert with the two Sox1 and Sox3 (Lefebvre *et al.* 2007). In addition, recent findings show that SoxC induces neuronal commitment of adult NSCs in hippocampal dentate gyrus (Mu *et al.* 2012). Therefore, as during embryogenesis, stage-specific genetic programs underlying the development of new hippocampal neurons seem coordinated by sequential acting of SoxB1 and SoxC transcription factors. Interestingly, continuum between embryonic and adulthood Sox functions appears to be present in several other tissues. For example, lens, bone and stomach continue to express Sox2 in some adult stem and progenitor cells (reviewed in Sarkar & Hochedlinger 2013). The study of the Sox family role during adulthood constitutes an opportunity to better understand molecular mechanisms underlying homeostasis, repair and regeneration in animals.

Box 1: Sox genes in human genetic and congenital disorders

The importance of the Sox gene family in human development implies that they are at the origin of several human disorders. Thus, mutations in the seven *SRY*, *SOX2*, *SOX3*, *SOX9*, *SOX10*, *SOX17* and *SOX18* genes are associated with human diseases in the Online Inheritance in Man (OMIM) database. As stated above, mutations in the *SRY* gene cause sex reversal in humans. This was the first *SOX* gene implicated in human disease in 1990 (Jager *et al.* 1990). During the 1990s, and especially the last ten years, researchers have described several other Sox genes in human disease. Thus, campomelic syndrome with autosomal sex reversal, that affects bone and connective tissues, is caused by mutations in *SOX9*, which lead to C-truncated proteins (Foster *et al.* 1994, for a detailed review see Lee & Saint-Jeannet 2011). Mutations in *SOX10* cause Waardenburg syndrome and Hirschsprung disease in humans and the equivalent dominant megacolon (Dom) phenotype in mice (Pingault *et al.* 1998, Southard-Smith *et al.* 1998). More recently, Fantes and collaborators show that mutations in *SOX2* cause microphthalmia, optic nerve hypoplasia and abnormalities of the central nervous system (Fantes *et al.* 2003). In the related *SOX3*, mutations cause X-linked mental retardation and

panhypopituitarism, in which patients present facial anomalies, low growth and infantile behavior (Laumonnier *et al.* 2002). Finally, vesicoureteral reflux and hypotrichosis-lymphedema-telangiectasia syndrome are caused by mutations in *SOX17* and its related *SOX18*, respectively (Gimelli *et al.* 2010, Irrthum *et al.* 2003).

Sox and reproductive biology

Sry gene: unique function during mammalian embryogenesis

The mammalian *Sry* gene was initially found by gene mapping to the sex-determining region of the human Y chromosome that caused sex reversal of XX males and XY females (Sinclair *et al.* 1990). Mutation in *SRY* in the genome of XY females and mutational analysis of *Sry* orthologue in mouse confirm that this gene initiates testis differentiation during embryogenesis (Hawkins 1993). This discovery was a starting point for investigation of the genetic program that led to male pathway in mammals. In contrast to the wide expression of the other Sox factors in time and space during development, *Sry* is transiently expressed in the mammalian embryos for a few hours in a single tissue, the gonadal somatic cells. During this short time, *SRY* targets a specific *Sox9* enhancer, and act in synergy with the orphan nuclear receptor SF1 (Steroidogenic factor 1), committing gonadal somatic cells into Sertoli cell fate (Sekido & Lovell-Badge 2008). *SRY* certainly proceed also as architectural protein, altering chromatin configuration, and activating male or repressing female specific genes (Brennan & Capel 2004). However, genetic pathway involved in testis differentiation in mammals is still elusive and will be subject to exciting next discoveries. Interestingly, *Sry* expression is conserved in many adult male mammalian tissues, particularly in the central nervous system, such as its closely related *SoxB1* members, suggesting a role not restricted to the sex determination (Turner *et al.* 2011). It is worth noting recent findings on epigenetic mechanism underlying temperature-dependent sex determination in the european sea bass suggest a potential role of Sox factors in this biological process (Navarro-Martin *et al.* 2011). These studies show the importance of the Sox family in the genetic and environmental (epigenetic) sex determination of vertebrates.

Sox factors in gonads

Master sex-determining genes, such as *Sry* in mammals, are highly variable in vertebrates (Schartl 2004). However, molecular mechanisms underlying sex differentiation are more conserved. Interestingly, the role of Sox factors in this biological process is ancient and conserved from fly to mammals. For instance, the *SoxE* in

Drosophila, named Sox100B, is required for somatic testis differentiation, such as its orthologue Sox9 in vertebrates (Schartl 2004, Nanda *et al.* 2009). In fish, the Sox factors are widely expressed in male and female gonads, highlighting their role in reproductive development (Santos *et al.* 2007). For instance, in the european sea bass, *sox17* and *sox19* present sex differences in expression, with up-regulation during ovarian differentiation (Navarro-Martin *et al.* 2009, Navarro-Martin *et al.* 2012). Otherwise, Zapater and collaborators recently shown that different Sox factors expressed in ovaries seem participated to the formation of the pelagic phenotype (buoyant) of marine teleost eggs. Indeed, the aquaporin water channel (*aqp1ab*), which has an essential role in oocyte hydration, may represent one of the targets of Sox3 associated with oocyte differentiation in teleosts (Zapater *et al.* 2013). Interestingly, Sox3 is required in the mouse for normal oocyte development and for male testis differentiation, suggesting a conserved role in the reproductive development of vertebrates (Weiss *et al.* 2003). The general message that has emerged from these studies is that Sox factors play important role in gonad development and more generally in reproductive biology, a major engine of evolution. For instance, Finn and Kristoffersen argue that the pelagic phenotype of marine teleost eggs were a crucial adaptation that lead to the successful radiation of teleostean fishes in the oceanic environment (Finn & Kristoffersen 2007).

Sox factors: neural crest, cell migration and vertebrate evolution

Co-option of Sox proteins into neural crest stem cells in vertebrates

The three SoxE factors, Sox8, Sox9 and Sox10, are fundamental for the development of the neural crest cells (NCCs), a cell population that emerges from the dorsal neural tube during early vertebrate embryogenesis. These remarkable cells, which are migratory and multipotent, differentiate into a very wide range of derivatives, divided into the two ectomesenchymal (such as bone and cartilage), and non-ectomesenchymal (neuron, glia and melanocytes) categories (Donoghue *et al.* 2008). The discovery that neural crest disorder in the Hirschsprung disease is associated with disrupted expression of *Sox10* initiated research on the role of Sox proteins in the neural crest development (Southard-Smith *et al.* 1998). Further works using Sox10-null mouse provide evidences for its essential role for non-ectomesenchymal derivatives differentiation (Dutton *et al.* 2001). Moreover, Sox9 factor is necessary for chondrocyte differentiation and cartilage formation (Bi *et al.* 1999). Interestingly, the role of SoxE in neural crest development seems to have been co-opted in stem vertebrates. Indeed, the single SoxE orthologue in the amphioxus is not expressed at the right place and at the right time during

embryogenesis (Meulemans & Bronner-Fraser 2004). Moreover, Sox10 replacement by the *drosophila* SoxE (named Sox100B) in mouse restore glial differentiation, suggesting properties already present in the fly orthologue (Cossais *et al.* 2010). The acquisition of neural crest feature in vertebrates, and its subtending gene-regulatory networks, is proposed to be a key factor during vertebrate evolution that facilitated the remodeling of the head (Gans and Northcutt, 1983).

Post-translational modifications of SOX proteins

Sox members paired off with different partners explain their multiple tissue-specific functions during embryonic development (Kamachi *et al.* 2000). Moreover, post-translational modifications (PTMs) of Sox proteins such as SUMOylation and phosphorylation mediate these changes in partner interactions. For instance, during neural crest development in vertebrates, PTMs of Sox9 is essential for the acquisition of migratory capabilities of NNCs. Indeed, extracellular signal induces the canonical Wntless-type (Wnt) pathway activation that promotes Sox9 phosphorylation, followed by its SUMOylation. These PTMs promote the interaction of Sox9 with the zinc-finger transcription factor Snail2, necessary for NCCs migration initiation (Liu *et al.* 2013). Thus, Sox9 promotes first NCCs identity during early neural crest development, and second, paired off with Snail2 to induce NNCs migration. This study provides new cues for the role of the Sox factors in delamination and in the acquisition of migratory property of neural crest cells. Moreover, given the similarities between neural crest migration and cancer invasion, it may be helpful to use NCCs as a model to better understand the role of Sox factors in cancer development, particularly in the metastasis mechanism (Theveneau & Mayor 2012).

Box 2 : Sox in cancers and metabolic diseases

In addition and in agreement with their critical functions in cell regulation and cell migration, Sox transcription factors are implicated in a wide range of cancers. Recent reviews show the importance of this gene family in cancers from multiple tissue types, such as colon, pancreas, breast, lung and brain (for detailed reviews, see Gracz & Magness 2011 and Vervoort *et al.* 2012). A deregulation of Sox factors is generally observed in tumors, that provokes an alteration in expression of downstream target genes, but their oncogenic role is still elusive. Interestingly, recent studies highlight the importance of pioneer factors during tumorigenesis by driving changes in transcriptional programs (Magnani *et al.* 2011). Such as the other pioneer factors PBX1,

AP-2 γ and FOXA1, Sox factors are generally over-expressed in cancer, and these changes can have consequences for the progression of tumors. Understanding the contribution of these pioneer factors during tumorigenesis should be promising for new prognostic markers or therapeutic targets development. Moreover, Sox factors could be implicated in metabolic diseases. For instance, Sox6 deregulation in pancreatic β -cells is associated to insulin-dependent disorders, such as type 2 diabetes and obesity (Gracz & Magness 2011).

Conclusion

Since the discovery of its first member in 1990, the Sox family has been subject to numerous studies, which revealed its ancient origin and prominent role in animal biology. The Sox family, which originated in the stem eukaryote, has expanded over time through duplication events, including whole genome duplications during vertebrate evolution. This expansion has been accompanied by the functional diversification of the different Sox members. Indeed, the characterization of hundred of Sox members in dozen of species highlighted their diverse roles in development, physiology and diseases. Moreover, many studies show the importance of the Sox factors in vertebrate evolution, especially in the emergence of neural crest cells, as well as in mammalian evolution with the fundamental role of SRY in sex. Altogether, these results explain the growing interest granted to the Sox family during the last decade. Further research on the Sox family, especially in their regulation and mode of action, should provide relevant answers in the three evolutionary, biological and medical perspectives.

- Bergsland, M. et al., 2011. Sequentially acting Sox transcription factors in neural lineage development. *Genes & development*, 25(23), p.2453–2464.
- Bi, W. et al., 1999. Sox9 is required for cartilage formation. *Nature genetics*, 22(1), p.85–89.
- Bowles, J., Schepers, G. & Koopman, P., 2000. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *DEVELOPMENTAL BIOLOGY*, 227(2), p.239–255.
- Brennan, J. & Capel, B., 2004. One tissue, two fates: molecular genetic events that underlie testis versus ovary development. *Nature Reviews Genetics*, 5(7), p.509–521.
- Cary, P.D. et al., 2001. Solution structure and backbone dynamics of the DNA-binding domain of mouse Sox-5. *Protein science: a publication of the Protein Society*, 10(1), p.83–98.
- Conant, G.C. & Wolfe, K.H., 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, 9(12), p.938–950.
- Coriat, A.M. et al., 1993. PCR amplification of SRY-related gene sequences reveals evolutionary conservation of the SRY-box motif. *Genome Research*, 2(3), p.218–222.
- Cossais, F. et al., 2010. Replacement of mouse Sox10 by the Drosophila ortholog Sox100B provides evidence for co-option of SoxE proteins into vertebrate-specific gene-regulatory networks through altered expression. *Developmental Biology*, 341(1), p.267–281.
- Crane-Robinson, C. et al., 1998. The energetics of HMG box interactions with DNA. Thermodynamic description of the box from mouse Sox-5. *Journal of molecular biology*, 281(4), p.705–717.
- Denny, P. et al., 1992. A conserved family of genes related to the testis determining gene, SRY. *Nucleic acids research*, 20(11), p.2887.
- Donoghue, P.C.J., Graham, A. & Kelsh, R.N., 2008. The origin and evolution of the neural crest. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 30(6), p.530–541.
- Dutton, K. et al., 2001. Zebrafish colourless encodes sox10 and specifies non-ectomesenchymal neural crest fates. *DEVELOPMENT*, 128(21), p.4113–4125.
- Fantes, J. et al., 2003. Mutations in SOX2 cause anophthalmia. *Nature Genetics*, 33(4), p.462–463.
- Finn, R.N. & Kristoffersen, B.A., 2007. Vertebrate Vitellogenin Gene Duplication in Relation to the « 3R Hypothesis »: Correlation to the Pelagic Egg and the Oceanic Radiation of Teleosts. *Plos One*, 2(1).
- Foster, J.W. et al., 1994. Campomelic dysplasia and autosomal sex reversal caused by mutations in an SRY-related gene. *Nature*, 372(6506), p.525–530.
- Gans, C. & Northcutt, R.G., 1983. Neural crest and the origin of vertebrates: a new head. *Science (New York, N.Y.)*, 220(4594), p.268–273.
- Gao, N. et al., 2013. Structural basis of human transcription factor Sry-related box 17 binding to DNA. *Protein and peptide letters*, 20(4), p.481–488.
- Gimelli, S. et al., 2010. Mutations in SOX17 are associated with congenital anomalies of the kidney and the urinary tract. *Human Mutation*, 31(12), p.1352–1359.
- Gracz, A.D. & Magness, S.T., 2011. Sry-box (Sox) transcription factors in gastrointestinal physiology and disease. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 300(4), p.G503–G515.
- Graves, J.A.M., 2006. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5), p.901–914.

- Gubbay, J. et al., 1990. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*, 346(6281), p.245–250.
- Guth, S. & Wegner, M., 2008. Having it both ways: Sox protein function between conservation and innovation. *CELLULAR AND MOLECULAR LIFE SCIENCES*, 65(19), p.3000–3018.
- Han, F. et al., 2010. Characterization, phylogeny, alternative splicing and expression of Sox30 gene. *Bmc Molecular Biology*, 11.
- Hawkins, J.R., 1993. Mutational analysis of SRY in XY females. *Human mutation*, 2(5), p.347–350.
- Hokamp, K., McLysaght, A. & Wolfe, K.H., 2003. The 2R hypothesis and the human genome sequence. *Journal of structural and functional genomics*, 3(1-4), p.95–110.
- Holland, L., 2009. Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *NATURE REVIEWS NEUROSCIENCE*, 10(10), p.736–746.
- Van Houte, L.P. et al., 1995. Solution structure of the sequence-specific HMG box of the lymphocyte transcriptional activator Sox-4. *The Journal of biological chemistry*, 270(51), p.30516–30524.
- Irrthum, A. et al., 2003. Mutations in the transcription factor gene SOX18 underlie recessive and dominant forms of hypotrichosis-lymphedema-telangiectasia. *American journal of human genetics*, 72(6), p. 1470–1478.
- Jager, M. et al., 2006. Expansion of the SOX gene family predated the emergence of the Bilateria. *Molecular phylogenetics and evolution*, 39(2), p.468–477.
- Jauch, R. et al., 2012. The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition. *The Biochemical journal*, 443(1), p.39–47.
- Kamachi, Y., Uchikawa, M. & Kondoh, H., 2000. Pairing SOX off: with partners in the regulation of embryonic development. *Trends in genetics: TIG*, 16(4), p.182–187.
- King, N. et al., 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, 451(7180), p.783–788.
- Larroux, C. et al., 2006. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evolution & Development*, 8(2), p.150–173.
- Larroux, C. et al., 2008. Genesis and Expansion of Metazoan Transcription Factor Gene Classes. *Molecular Biology and Evolution*, 25(5), p.980–996.
- Laumonnier, F. et al., 2002. Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. *American journal of human genetics*, 71(6), p.1450–1455.
- Lee, Y.-H. & Saint-Jeannet, J.-P., 2011. Sox9 function in craniofacial development and disease. *genesis*, 49(4), p.200–208.
- Lefebvre, V. et al., 2007. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *INTERNATIONAL JOURNAL OF BIOCHEMISTRY & CELL BIOLOGY*, 39(12), p. 2195–2214.
- Liber, D. et al., 2010. Epigenetic Priming of a Pre-B Cell-Specific Enhancer through Binding of Sox2 and Foxd3 at the ESC Stage. *Cell Stem Cell*, 7(1), p.114–126.
- Liu, J.A.J. et al., 2013. Phosphorylation of Sox9 is required for neural crest delamination and is regulated downstream of BMP and canonical Wnt signaling. *Proceedings of the National Academy of Sciences*, 110(8), p.2882–2887.

- Magnani, L., Eeckhoute, J. & Lupien, M., 2011. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics*, 27(11), p.465–474.
- Malki, S., Boizet-Bonhoure, B. & Poulat, F., 2010. Shuttling of SOX proteins. *The international journal of biochemistry & cell biology*, 42(3), p.411–416.
- Meulemans, D. & Bronner-Fraser, M., 2007a. Insights from amphioxus into the evolution of vertebrate cartilage. *PloS one*, 2(8), p.e787.
- Meulemans, D. & Bronner-Fraser, M., 2007b. The amphioxus SoxB family: implications for the evolution of vertebrate placodes. *International journal of biological sciences*, 3(6), p.356–364.
- Mu, L. et al., 2012. SoxC Transcription Factors Are Required for Neuronal Differentiation in Adult Hippocampal Neurogenesis. *The Journal of Neuroscience*, 32(9), p.3067–3080.
- Nanda, S. et al., 2009. Sox100B, a Drosophila Group E Sox-domain Gene, Is Required for Somatic Testis Differentiation. *Sexual Development*, 3(1), p.26–37.
- Navarro-Martín, L. et al., 2012. Characterisation and expression during sex differentiation of Sox19 from the sea bass *Dicentrarchus labrax*. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology*, 163(3-4), p.316–323.
- Navarro-Martín, L. et al., 2009. Different sox17 transcripts during sex differentiation in sea bass, *Dicentrarchus labrax*. *Molecular and cellular endocrinology*, 299(2), p.240–251.
- Navarro-Martín, L. et al., 2011. DNA Methylation of the Gonadal Aromatase (cyp19a) Promoter Is Involved in Temperature-Dependent Sex Ratio Shifts in the European Sea Bass. *PLoS Genet*, 7(12), p.e1002447.
- Ner, S.S., 1992. HMGs everywhere. *Current biology: CB*, 2(4), p.208–210.
- Okuda, Y. et al., 2006. Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. *Developmental Dynamics*, 235(3), p.811–825.
- Palasingam, P. et al., 2009. The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *Journal of molecular biology*, 388(3), p.619–630.
- Paps, J., Holland, P.W.H. & Shimeld, S.M., 2012. A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Briefings in functional genomics*, 11(2), p.177–186.
- Phochanukul, N. & Russell, S., 2010. No backbone but lots of Sox: Invertebrate Sox genes. *INTERNATIONAL JOURNAL OF BIOCHEMISTRY & CELL BIOLOGY*, 42(3), p.453–464.
- Pingault, V. et al., 1998. SOX10 mutations in patients with Waardenburg-Hirschsprung disease. *Nature Genetics*, 18(2), p.171–173.
- Reményi, A. et al., 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes & development*, 17(16), p.2048–2059.
- Santos, E.M. et al., 2007. Molecular basis of sex and reproductive status in breeding zebrafish. *Physiological Genomics*, 30(2), p.111–122.
- Sarkar, A. & Hochedlinger, K., 2013. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell stem cell*, 12(1), p.15–30.
- Schartl, M., 2004. Sex chromosome evolution in non-mammalian vertebrates. *Current Opinion in Genetics & Development*, 14(6), p.634–641.
- Sekido, R. & Lovell-Badge, R., 2008. Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. *Nature*, 453(7197), p.930–934.

- Sinclair, A.H. et al., 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, 346(6281), p.240–244.
- Southard-Smith, E.M., Kos, L. & Pavan, W.J., 1998. Sox10 mutation disrupts neural crest development in Dom Hirschsprung mouse model. *Nature genetics*, 18(1), p.60–64.
- Theveneau, E. & Mayor, R., 2012. Neural crest delamination and migration: From epithelium-to-mesenchyme transition to collective cell migration. *Developmental Biology*, 366(1), p.34–54.
- Turner, M.E. et al., 2011. Sry, more than testis determination? *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 301(3), p.R561–R571.
- Vervoort, S.J., van Boxtel, R. & Coffey, P.J., 2012. The role of SRY-related HMG box transcription factor 4 (SOX4) in tumorigenesis and metastasis: friend or foe? *Oncogene*, 32, 3397-3409.
- Weiss, J. et al., 2003. Sox3 is required for gonadal function, but not sex determination, in males and females. *Molecular and cellular biology*, 23(22), p.8084–8091.
- Werner, M.H. et al., 1995. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell*, 81(5), p.705–714.
- Zapater, C. et al., 2013. Primary oocyte transcriptional activation of aqp1ab by the nuclear progesterone receptor determines the pelagic egg phenotype of marine teleosts. *Developmental Biology*, 377(2), p. 345–362.
- Zhao, C., Deng, W. & Gage, F.H., 2008. Mechanisms and Functional Implications of Adult Neurogenesis. *Cell*, 132(4), p.645–660.

RESUME

Les duplications de gènes et de génome sont considérées comme des moteurs de l'évolution des génomes eucaryotes. Trois duplications de génome complet (ou polyploïdisations) sont survenues au cours de l'évolution des vertébrés, dont deux à la base des vertébrés, et une troisième chez l'ancêtre commun des poissons téléostéens. La diversité morphologique, anatomique et écologique des espèces qui partagent un ancêtre commun polyploïde chez les chordés suggère un rôle des duplications de génome dans la diversification des espèces. En particulier, les duplications de génome semblent avoir facilité l'émergence du plan d'organisation des vertébrés, et être à l'origine de la radiation évolutive survenue chez les poissons téléostéens. Cependant, la portée évolutive des duplications de génome, et notamment les deux hypothèses majeures formulées ci-avant, restent des questions ouvertes et en grande partie non résolues. Le groupe des téléostéens, qui compte plus de la moitié des espèces vertébrés existantes et partage un ancêtre commun polyploïde, constitue un modèle pertinent pour évaluer la contribution des duplications de génome dans l'expansion des familles multigéniques chez les vertébrés, pour comprendre les mécanismes évolutifs qui façonnent l'évolution des familles de gènes, et finalement tester les hypothèses moléculaires qui peuvent relier duplication de génome et biodiversité. Ainsi, nous avons étudié l'impact de la duplication de génome survenue à la base des téléostéens sur l'évolution de la famille multigénique *sox*, essentielle pour le développement et l'homéostasie des vertébrés. Notre analyse du contenu et de l'organisation des gènes *sox* dans 15 génomes de vertébrés, dont 10 téléostéens, révèle une importante expansion de l'ensemble de la famille des gènes *sox* dans ce vaste groupe de vertébrés, et démontre que cette expansion est essentiellement due à la duplication de génome survenue à la base des téléostéens. Les gènes *sox* dupliqués par duplication de génome semblent avoir été perdus par non-fonctionnalisation dans certaines lignées, et préservés en deux copies par sous-fonctionnalisation et/ou néo-fonctionnalisation dans certaines autres lignées. Notre étude indique en effet une divergence lignée-spécifique des patrons d'expression entre les gènes *sox* dupliqués chez différentes espèces de téléostéens. Ainsi, l'expansion du répertoire des gènes *sox* à la base des téléostéens semble avoir été suivi d'une évolution lignée-spécifique du contenu et des fonctions de la famille des gènes *sox* chez les poissons téléostéens. Cette étude supporte l'hypothèse d'un rôle des duplications de génome dans l'enrichissement et la diversification subséquente des répertoires de gènes du développement tels que les gènes *sox*, et son rôle potentiel dans la diversification des espèces vertébrés.

Mots-clés :

Duplication de génome, famille Sox, poissons téléostéens, biodiversité